

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

Corresponding Author: Dr Adam Brandt

Any redactions in this file are there to maintain patient confidentiality, the confidentiality of unpublished data, or to remove third-party material.

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Referee #1

(Remarks to the Author)

General comments

The authors address a key issue surrounding the use of point-source satellite methane detection and quantification – the lack of large-scale independent, controlled release testing. The authors conducted a year-long single-blind release study testing both different satellite sensors and different retrieval algorithms from different participants. The paper examines detection and quantification performance and discusses reasons for differing performance across 3 broad types of satellite instrument.

This is an important topic with broad reach beyond the satellite retrieval community. There is a lot of public interest in the use of point-source satellite data and I have encountered two extremes outside of academia: those who see it as a panacea for all evils and those who dismiss it as highly inaccurate and functionally useless. Fortunately, this work presents an objective evidence base to inform a more nuanced interpretation of the merits and limitations.

The scale of this work should not be underestimated. It is a highly impressive accomplishment to manage so many releases across a full year and to coordinate across the overpass times of numerous satellites and the number of participating quantification teams.

By and large the methods and paper structure follows a well-established protocol from the same group outlining the release approach, an evaluation of detects and non-detects and a quantification parity plot for each satellite. The paper concludes that satellite measurements can allow accurate detection and unbiased quantification.

However, there are a long list of caveats to that statement, and if there's one criticism it's that these are not more front and center throughout. Given the significant investment and interest in satellite plume detectors, it is easy to see how the concluding statements may start to runaway from the realms of scientific nuance like an unstoppable freight-train on the captive eyes of the unsuspecting stakeholder.

Examples of these caveats include the still relatively small sample sizes for individual satellites (despite the impressive scale of the experiment), sanitized testing conditions and teams able to run at higher sensitivity modes than they normally would. These are discussed by the authors, but don't make it into the abstract or some of the more sweeping statements within the paper. The other main point is to emphasize in the broad concluding statements is that these findings are specific to plume detectors and the findings relate only to the detection of large individual plumes, not a finding on the ability of satellites to accurately quantify emissions more broadly.

Some presentation of results may be inadvertently open to misinterpretation – for instance a 100% TP rate within a 0-250kg/h bracket may give the false impression that a particular offering can cover this full range with 100% accuracy. The reality is that the sample size is small, most of the testing in this bracket appears to have been conducted above 100 kg/h and the smallest non-zero release may have been around 50 kg/h. The danger is that the reader is unaware of the fine print and walks away with an overly optimistic outlook.

Perhaps the most useful insights are the Group C as they have relatively good coverage across the year and the higher detection limits ensure a far greater proportion of testing occurred within the a priori marginal detection range. But curiously, the probability of detection is not investigated in any detail.

The lack of involvement of the participating quantification teams in the final analysis is a missed opportunity. There is no sense of scientific improvement or endeavor towards some greater sense of verisimilitude. It's clear that subjective decisions can greatly influence estimated emissions from the open data satellites such as EMIT and Sentinel-2. Two teams presented with the same data can arrive at very different results. But the paper never really gets at why. Can this paper point towards ways to reduce the reliance on subjectivity?

None of this should take away from the importance and quality of the scientific work. Merely the implications of its publication perhaps warrant a little more oversight of its messaging.

Specific comments

Abstract – Shouldn't the abstract mention the range across which the detection statistics apply? Because they are specific to the distribution of release magnitudes of this test.

Abstract L17 – What does 15 times larger mean in this context? Release size, number of releases or number of participants?

Abstract L27-28: "Broad land scanners performed poorly but have consistent global coverage and vast historical archives, motivating additional work". This needs further clarification – their performance is worse for detection undoubtedly. But for those plumes that they do detect, the quantification accuracy doesn't appear to drop off markedly compared to Group A and Group B.

P2. L67 – Isn't IMEO part of UNEP rather than the UN?

P5. L140 – Is IMEO a research institution?

P5. L154-156 – Is there any significance to the phasing of the experiments across the year? Did any team update their algorithm based on the ground truth data of a previous phase and improve in subsequent phases?

P10. L227 – "These FPS can be more or less impactful". Suggest this point could be more forceful. FPs are the fastest way to burn credibility.

P11. L245 – Does this have to be sidelined into future work? Even for this one example, wouldn't this paper offer so much more if it investigated why the same satellite retrieval can result in such different solutions? Investigator decisions are clearly important with these indirect forms of measurement. How can this work point to ways to reduce the impacts of subjective investigator decisions?

P17. L375 - This bin of 200 kg/h is too open ended and gives the wrong impression. Most of these satellites were not tested below 50 kg/h and there is clearly a lower bound above 0 on which this statement might apply.

P17. L 377 – Is there a citation for this GHGSat Permian evaluation? Not clear how this 50% POD detection limit was arrived at.

P17. L375-389 – This is a long list of important caveats. And ultimately it would be very hard to account for every one through controlled release testing. Does this point to natural limits on the use of release testing? Potentially statistics such as probability of detections may need to be derived via other means where the ground-truth is not necessarily controlled but larger sample sizes can be collected. This is somewhat implied in the discussion of Tanager detection distribution in Section S3.6.

P17. L392-393 – The theoretical detection limit is estimated at 1000 kg/h for S2. Perhaps it needs mentioning that 303 of 345 detection outcomes (88%) took place below this simulated detection limit. In other words, the poorer detection performance is entirely expected based on a priori knowledge. I assume this is a much higher percentage than Groups A and B. Does this affect the interpretation of statistics across groups? For comparison what percentage of testing took place below the simulated (or empirically derived) detection limits of the other satellite groups? Assuming this is small, is this why the paper does not explore probability of detection curves? For plume detecting satellites to be used for quantification purposes across broader timescales don't we need more empirical data on this?

P18. L405 – The relationship between the strictness of detection criteria and the application of the quantification team seems quite pertinent. It's noticeable that commercial participants tend to set more lenient criteria (resulting in more FPs) than say an operational alert program. FPs are the fastest way to burn credibility and have the potential to jeopardize methane mitigation initiatives. Perhaps this point needs to be made more strongly.

P19. L415 – Any reason to pick out Nigeria and Venezuela from oil-producing tropical countries? Are they particularly

plume-heavy relative to other tropical producers?

P18. L417 – In this experiment participants have the advantage of knowing the source location. This is akin to an operational situation where the target is known and tasked specifically. But for the other operational case of unknown unknowns would participants achieve successful detections in >45% cloud cover without this knowledge?

P19. L426-429 – The point about plume detecting satellites being important in basins where sources >100kg/hr dominate is fine, but is the Permian a good example? For instance, Williams et al (2025) arrived at an opposite distribution where 70% emissions are below 100 kg/h rather than above. Appreciate this is an area of debate but given that, it might be more appropriate to keep the point abstract rather than specific.

P19. L434-436 – Does it really matter whether a controlled release targeting satellites can cover this, that and every other detail? There will always be aspects of real life that can't be controlled. And should we try? The scale of this experiment is mightily impressive and yet the returned sample sizes remain relatively small for each individual satellite because of revisit times and quality control. As mentioned above, perhaps the solution is to sacrifice near-perfect emissions knowledge for larger statistical sample sizes.

P19. L441 – It is hard to agree with the characterization of the emission rates tested by Group A in this study as “relatively small”. They could easily be called “exceedingly large”. The majority are above 100 kg/h, commonly used as the arbitrary super-emitter definition. And as Fig S4 shows, the real-world detection performance certainly drops off below 100 kg/h.

Supplement Table S3 – A number of teams used HRRR wind data – a product only available for CONUS. What would those teams use in other parts of the world and is there any indication of how this model compares to global reanalysis products such as ERA5 and thus whether similar performance would be expected for equivalent arid sites in other parts of the world?

Table S3 - ERDDAP is a data server containing multiple datasets rather than a reanalysis product – what actual dataset was used?

Table S3 – Is OpenWeatherMap a unique reanalysis product or a commercial data server hosting specific reanalysis datasets?

References

Williams et al. (2025) Small emission sources in aggregate disproportionately account for a large majority of total methane emissions from the US oil and gas sector, *Atmos. Chem. Phys.*, 25, 1513–1532, <https://doi.org/10.5194/acp-25-1513-2025>

Referee #2

(Remarks to the Author)

The authors present the results of a year-long single-blind controlled release experiment evaluating satellite observations of methane point source plumes. The experiment evaluates 11 satellite systems, with data for hundreds of controlled releases (and zero-emission blanks) analyzed by 15 satellite teams from a variety of institutions and countries. This is by far the largest-scale controlled release trial for satellites to-date and a very important study for the methane remote sensing and mitigation communities. The key finding is that current satellite technologies work, but with a hierarchy of performance. Purpose-built methane sensors show excellent detection/quantification capabilities, while the performance of more general/repurposed instruments is at best comparable but overall moderately degraded, and that of global land imagers with very coarse spectral resolution but wide spatiotemporal coverage is more severely degraded. The experimental design, data analysis methodologies, presentation of results, and writing are all of excellent quality and I support publication of the study in *Nature*.

I have one “major” comment and several minor ones that I would like the authors to address:

My major comment is a suggestion. I would strongly encourage the authors to perform one additional experiment, which is to evaluate the performance of the data providers as a “crowd” for plumes analyzed by multiple teams. The paper does a very good job describing individual team/sensor performance and summarizing overall capabilities of the data provider community. But is there a “wisdom of crowds” effect to document here? In other words, when analyzing the same plume, does the ensemble estimate (mean, median, something else) from a collection of data providers outperform that from any one data provider? This would be very good to know, as it could temper some of the concerns around inter-provider variability that have been recently highlighted by the methane remote sensing user/provider communities (and highlighted in the discussion of Figure 3, where the provider mean emission rate estimate is 628 kg/h – 2x closer to the truth than the individual team estimates on average).

Minor comments:

Line 21: “Across all sensors and teams, 70.6% correctly identified the presence or absence of emissions.” I initially read the percentage as referring to the sensors and teams, but that’s obviously wrong – it refers to the >2250 estimates. Rephrase for clarity?

Line 27: “Broad land scanners performed poorly but have consistent global coverage and vast historical archives, motivating

additional work.” Yes, but as mentioned several times in the manuscript, this is to be expected given the relatively low release rates (compared to land scanner detection limits). I agree wholeheartedly with the statement that follows about the motivation to improve land scanner retrievals, but also wonder whether the detection limit caveat can be highlighted here.

Line 31: “Our findings show that high-quality methane measurements from space can allow accurate detection and unbiased quantification, initiating a new era in international greenhouse gas mitigation.” I believe this is an overstatement and should be dialed back / rephrased. This is a very important study that can serve as a key reference for the reliability of satellite methane point source monitoring, but it does not “initiate a new era” in GHG mitigation. Satellite-based mitigation has been ongoing for quite some time already.

Line 47: “These instruments range from multispectral government satellites with broad Earth-observing missions (e.g. Sentinel-2) to commercial constellations of multiple satellites that offer GHG monitoring as a service.” The ESA Sentinel-2 mission overview page is cited here, which is a bit odd since it says nothing of methane. Something like the Jacob 2022 review or the Varon 2021 Sentinel-2 methane paper might be a better citation.

Line 52: “These are the first systems that are both designed specifically to detect methane and are committed to releasing large-scale datasets into the public domain.” A minor correction, but this statement is not really accurate (as becomes clear later in the sentence). Other instruments, like TROPOMI, have been “designed specifically to detect methane” and release “large-scale datasets into the public domain.” Carbon Mapper and MethaneSAT were the first to do this for methane point sources specifically (to my knowledge) and could also be differentiated as non-profit methane satellite missions.

Line 110: “(6) a complete year of data, allowing for more variety in weather conditions and sun angles.” I was interested to see analysis of performance by sun angle / weather conditions, but aside from cloud cover, I don’t believe this ever appears in the manuscript. Perhaps I missed it, but if not then I would suggest clarifying this 6th contribution.

Line 131: “SRON (Space Research Organization Netherlands)”. Is this acronym intentionally reversed?

Line 169: “EMIT overpass frequency is irregular because it is mounted on the ISS and has an unpredictable non-polar orbit.” The orbit of the ISS is not unpredictable, but its precession does produce irregular revisit rates. Please rephrase for clarity.

Line 196: “This should not be taken as evidence of the general ability of Group C instruments to see very large (>1500 kg/h) emission sources, as our equipment can only release volumes at the expected low end of the detection range for these instruments.” Is there a good reference to make the point that the Group C instruments do routinely detect large sources globally? Perhaps something from IMEO’s MARS work could make that point.

Line 209: “Rates must therefore be adjusted for cloud coverage in a given location.” I do not understand this. Is it saying that the TP rate would be degraded for partially cloudy scenes? How would you adjust the rates to account for cloud cover? Please elaborate/rephrase to clarify.

Line 217: “In contrast, outcomes for Sentinel-2 and Landsat varied substantially among teams. Out of 71 non-zero releases for Sentinel-2, two teams made 0 or 1 TP detections while two other teams made >35 TP detections (see SI Figure S2 for each team’s outcomes by release rate); Sentinel-2 F1 scores range 0.0-0.73 (SI Table S6).” I thought this was a very interesting finding and was disappointed not to get some kind of “concluding sentence” afterwards, like the preceding one stating that “the very similar TP rate” despite the “heterogeneity of analysis approaches” for Group B instruments shows that their detection limits are governed by spectral information. Does the much wider range of results for the land scanners mean that there is potentially more information available from e.g. Sentinel-2 than is accessible by standard data analysis methods, and that AI (used by some teams and not others) can unlock? Does it mean something else? What did the high-performing Group-C teams do better than the low-performing teams?

Line 241: “two stand out as having significantly different smoothing and brightness.” Presumably those are Google and Orbis Earth? Would be worth making explicit.

Line 244: “There is no apparent trend between errors in assumed windspeed and enhancement concentration in this example.” It was not clear to me whether this means that the teams all used similar wind speeds to estimate the emissions. Is that the correct interpretation? Or is the sentence saying that windspeed errors and methane plume concentrations are not correlated (why would they be)? Using different wind speeds could of course lead to large differences, as could different masking techniques. Please clarify this sentence.

Figure 5: On the inclusion (or not) of FP in the emission rate regressions, I see in the supplement that both approaches were tested. But I don’t see any summary of results for the TP-only regressions. If I have not missed it, would it be feasible to include in some form in the supplement? Some of the default FP+TP regression lines going in unexpected directions are not very informative.

Line 338: “Quantification was also largely unbiased. For example, Carbon Mapper successfully detected every non-filtered release (n=31) with Tanager-1 in the year-long experiment. They also reported zero false positives across five no-release controls.” The flow of ideas here, from the first to second and third sentences quoted, is not very clear to me.

Line 378: “A recent evaluation comparing GHGSat to airborne instruments in the Permian Basin found that its 50% POD threshold is near 250 kg CH₄/h.” Please provide the citation for this evaluation. The recent Jervis 2026 paper in ES&T could

be useful here – they show that field or "survey" detection probability is considerably worse than in controlled conditions: <https://pubs.acs.org/doi/abs/10.1021/acs.est.5c17089>

Referee #3

(Remarks to the Author)

A.Summary of key results

This one-year experiment ran from 1 January 2025 to 31 December 2025, consisting of 746 observational events including 481 quantitative controlled methane release tests and 265 blank controls. Twelve on-orbit hyperspectral and multispectral satellites were arranged to observe the ground-based controlled methane release activities. Fifteen data analysis teams from nine countries produced more than 2,250 emission estimates in total. The paper analyzes and compares the true positive rates, false positive rates, biases of methane releases detection from various satellite spectral imaging sensors. It also evaluates the impacts of cloud cover on methane detection performance and compares inversion approaches driven by artificial intelligence.

B.Originality and significance

This study provides an abundant dataset of ground-based controlled methane release experiments conducted in the desert area of Casa Grande, Arizona, USA, including blank observations with zero methane emissions. The tested methane emission rates range from 9.5 to 1,183.7 kg CH₄/h, and the dataset records diverse observational conditions such as albedo, solar elevation angle, wind speed and cloud coverage. This dataset delivers substantial value to research on satellite-based methane emission detection.

Furthermore, the comprehensive detection results covering different satellites, wind speed bins, cloud coverage, emission magnitudes and independent analysis teams carry important academic and practical implications for understanding methane monitoring performance of spaceborne spectral imagers.

Nevertheless, several points deserve clarification:

Four core components constitute the whole research workflow:

- ① Ground-based controlled methane release setup and concurrent meteorological measurements;
- ② Satellite overpass observations synchronized with pre-designed release/non-release test scenarios;
- ③ Over 2,250 release or non-release estimates generated from experimental data;
- ④ Integrated summary and analysis of all experimental setup and observational outcomes.

Components ② and ③ represent the core contributions from satellite operators and especially the 15 independent analysis teams. Their technical and academic inputs are indispensable to the valuable dataset and meaningful detection results presented in this paper. Although key participants are acknowledged at the end of the manuscript, the current author affiliation list does not fully capture the full scope of cross-team contributions to this large-scale collaborative field campaign.

C. Data & methodology

(1) Lines 50–55 state that analytical teams were blinded to all ground-truth data during each satellite overpass, including the occurrence of methane releases, exact release timings, emission flow rates and in-situ meteorological observations. While concealing ground-truth emission data aligns with single-blind experimental protocols, wind speed measurements should arguably have been shared with participating teams. As illustrated in Supplementary Information Figure S7 (Line 596), wind speeds retrieved independently by individual teams deviate substantially from in-situ observations. Given that wind field characteristics dominate methane plume dispersion and quantification characteristics, such large wind estimation discrepancies inevitably undermine the reliability of cross-satellite detection performance evaluations.

This study aims to quantify the native detection performance of spaceborne methane sensors, and wind speed cannot be directly retrieved by any of the satellites assessed here. Inter-team wind speed estimation errors can reach multiple-fold magnitudes, introducing unquantifiable confounding biases into sensor performance assessments and derived statistical results. Wind field retrieval represents an independent research domain requiring dedicated methodological development and validation, and should not act as an uncontrolled confounding variable in inter-satellite comparative experiments.

(2) Table 1 contains an inaccurate parameter for the SWIR band of Gaofen-5: its spectral resolution is better than 10 nm (instead of 5 nm). The reference below can be cited for correction:

Li F, Lin K, Yan Y, et al. Comparing the performance of different hyperspectral satellite imaging spectroscopy in mapping methane point-source emissions. *Remote Sensing of Environment*, 2026, 334: 115224.

(3) Figure 1 indicates limited observational coverage for several platforms: Gaofen-5 and Ziyuan-1 each only acquired one release event analyzed by a single team; GHGSat, GHOSat and WorldView-3 also received analysis from merely one team each; Tanager-1 data was processed by two teams only. All remaining satellites had at least five independent analysis teams involved, while MethaneSAT provided no analytical results.

Coordinating satellite overpass schedules to align with ground release events and organizing multiple teams to process datasets from diverse satellites undoubtedly poses logistical challenges. However, this is a large-scale one-year campaign with 481 valid methane release tests, intended for publication in a high-impact journal such as *Nature*. Severe under-sampling for certain satellites and highly uneven distribution of analysis teams significantly compromise data completeness and methodological robustness, two critical standards expected of landmark interdisciplinary research designed to integrate cross-institutional resources to address global research gaps.

The Gaofen-5 hyperspectral payload shares a 30 m spatial resolution with Tanager-1 yet delivers a 60 km swath width—over three times wider than Tanager-1. A complete performance characterization of Gaofen-5 would greatly advance the development and application of this class of hyperspectral methane sensors, and the absence of sufficient Gaofen-5 evaluation constitutes a notable limitation of this otherwise valuable work.

(4) Figure S2 in Supplementary Information Line 466 shows that the two Tanager-1 analysis teams achieved perfect true positive detection across all release events. Similarly, certain teams working with EMIT, EnMAP and PRISMA also successfully detected every methane plume without missed signals. This creates ambiguity when interpreting overall performance statistics: it remains unclear whether cross-team disparities in detection outcomes stem primarily from analytical algorithms or intrinsic satellite sensor characteristics, which impairs the precision of inter-platform evaluation.

Figure 3 at Line 250 further demonstrates extreme variability in reconstructed methane plumes derived from ten independent teams processing identical EMIT overpass data.

The paragraph “Variable quantification performance across instrument classes” (Lines 284–291) reports that Group A sensors deliver slightly superior methane quantification precision compared with Group B, yet Group B features far narrower confidence intervals (Group A: 43.5% [95% CI: 34.9, 52.2]; Group B: 45.8% [95% CI: 42.0, 49.7]). These contrasting metrics raise questions about the extent to which cross-team algorithmic choices (alongside wind retrieval biases mentioned above) drive reported performance rather than inherent sensor performance.

Additionally, Lines 276–281 claim that detection results from Group A and Group C sensors show low sensitivity to wind speed, whereas Group B sensors exhibit strong wind dependence. This observation appears contradictory to the narrower confidence intervals measured for Group B instruments.

(5) Lines 260–270 report weaker correlation between cloud coverage and detection performance for Group A sensors relative to Groups B and C. Figure 4 presents counterintuitive patterns: Group A sensors produce false negatives under near-clear sky conditions, yet exclusively true positives at moderate cloud cover, with zero valid retrievals under 20–40% cloud fraction. For Group B sensors, extensive false positive and false negative detections occur below 40% cloud cover, while zero misdetections are recorded across 11 test cases at 40–60% cloud fraction. These inconsistent patterns highlight large uncertainties embedded within the evaluation workflow. Beyond ground-based cloud monitoring, all analysis teams should implement scene-level cloud masking algorithms using satellite imagery itself as a standard preprocessing step.

(6) Figure 5 shows that WorldView-3 achieves highly accurate retrievals for low-magnitude methane releases, yet error magnitudes rise substantially for large emission sources; Tanager-1 displays identical behavior. The manuscript should elaborate on the root causes of this trend. One potential contributing factor is the uniform, high-albedo desert surface of the test site. Supplementary Figures S76 and S77 further reveal that the ground release infrastructure occupies a large footprint with strong surface brightness, which likely exerts disproportionate interference on high-spatial-resolution payloads. The highly specialized, homogeneous terrain of this experimental site limits the generalizability of detection performance metrics to real-world methane emission landscapes.

D. Appropriate use of statistics and treatment of uncertainties

No comments.

E. Conclusions

The general conclusions agree with existing consensus on satellite methane sensing, while detailed quantitative results rely heavily on dataset quality and methodological limitations discussed in Section C (Data & methodology).

F. Suggested improvements

All revision recommendations are detailed within Section C (Data & methodology).

G. References

No comments.

H. Clarity and context

(1) Lines 50–60 compare the payloads of Tanager-1 and MethaneSat against TROPOMI. However, multiple hyperspectral sensors with 30 m spatial resolution were launched well before 2024, including Gaofen-5 (2018), PRISMA (2019) and EnMAP (2022). All of these platforms deliver drastically improved performance for methane point-source detection relative to TROPOMI. Therefore, a more relevant benchmark for Tanager-1 would be these peer 30 m hyperspectral instruments, and the manuscript should be revised accordingly.

Version 1:

Reviewer comments:

Referee #1

(Remarks to the Author)

Thanks for considering and addressing my comments. This is a quality study of high importance and impressive scale. Well

done.

(Remarks on code availability)

Referee #2

(Remarks to the Author)

I commend the authors for their thoughtful responses to the reviewer comments. Their manuscript is much improved and I recommend it be accepted for publication in Nature.

(Remarks on code availability)

Referee #3

(Remarks to the Author)

I fully appreciate the substantial effort the authors have devoted to addressing the reviewers' comments, with extensive additions and revisions made to the manuscript and supplementary materials that have markedly improved the overall quality of the paper. Nevertheless, as acknowledged by both the reviewers and the authors, this study involves numerous satellites and multiple research teams. While these are inherent features of the work, they also introduce intrinsic limitations: limited and sparsely distributed samples. Furthermore, algorithms adopted by different teams differ considerably, whereas cross-participation among research teams across different satellites remains extremely low, with only one or two teams performing analyses for some satellites. This inevitably amplifies uncertainties in the results and fails to reach the level of adequate inter-team cross-statistical validation expected for a single-blind test. In addition, methane plumes and emissions are subject to a wide range of influencing factors, including surface conditions, facility properties, wind speed, cloud cover, and solar-viewing geometry. All of the above factors create substantial complexity both for the authors in presenting their findings and for readers in interpreting the conclusions.

While reviewing this manuscript, I am impressed by the rich and diverse dataset and results presented in this work, yet I am left with substantial conceptual ambiguities following the concerns outlined above. Even for satellites analysed by a relatively large number of groups such as EnMAP, PRISMA and EMIT, the test results show considerable dispersion: some deviate far from the ideal 1:1 reference line, while others align closely with it. This creates significant confusion during my review process. It remains unclear whether better outcomes would be achieved with more extensive inter-team cross-validation or, conversely, with fewer participating teams. As noted by other reviewers and the authors themselves, owing to complexity, sample sparsity and other confounding factors, questions remain regarding the reliability of derived metrics such as minimum detection limits. There are open questions concerning the real-world detection capability of various satellite systems. For instance, WorldView-3 yields a striking minimum detection value of 9.5 kg/h — does this indicate that it delivers the best detection performance? This raises a fundamental question: is a single-blind experimental design inherently ill-suited for multi-target, multi-satellite campaigns that demand long-term execution and heavy logistical support? As the authors mention, it is difficult to achieve sufficient cross-comparison across satellites and research teams and to build a large, dense sample dataset. Without such data, statistical outputs cannot converge toward ground truth, which weakens the clarity and confidence of final conclusions.

Regarding the Main section at the beginning of the manuscript: based on the overall content and conclusions, this study is not primarily focused on Tanager1 and MethaneSat; indeed, MethaneSat ultimately provided no usable data for this work. In essence, this paper reports a comprehensive single-blind methane detection benchmark covering currently operational ultrahigh-spectral-resolution satellites (spectral resolution < 1 nm), high-spectral-resolution satellites (spectral resolution > 1 nm), high-spatial-resolution satellites, and wide-swath multispectral satellites. This feature should be explicitly stated in the text.

Moreover, I suggest a more objective description of the international development timeline for hyperspectral methane monitoring satellites, from Gaofen-5 (2018), PRISMA (2019), EnMAP (2022), EMIT (2022), through to Tanager1 (2024). These hyperspectral instruments belong to fundamentally the same category of payloads; therefore, the term "stepchange" may be inaccurate.

Following the launch of Gaofen-5, Ziyuan-1 and PRISMA, dozens of international research groups conducted methane detection campaigns over the Permian Basin in the United States. These studies first pushed the detection limit for point-source methane emissions down to the hundred-sof-kilograms-per-hour level. Key findings were originally reported in Science Advances (Itziar Irakulis-Loitxate et al., 2021). Strictly speaking, these payloads were not "repurposed towards methane observation".

Concerning the scarcity of Gaofen-5 and Ziyuan-1 data obtained in this experiment: the authors attribute this mainly to "Logistical & geopolitical barriers". I have consulted the operational authority for Gaofen-5, which demonstrates full willingness to participate in and support this valuable scientific experiment. Substantial funding and manpower from China have been invested in Gaofen-5 and Ziyuan-1 precisely to address these major scientific challenges facing human society. Admittedly, unfamiliarity and differing team-selection practices across national/regional collaborations have created practical communication and cooperation difficulties.

For Figure S1 in the new supplementary materials: I recommend removing the "min. test rate" column. The minimum detection limit value of 9.5 kg/h derived from WorldView-3 reaches performance levels comparable to airborne instruments. Yet the authors acknowledge in their response that such figures may be biased by high surface reflectance and other artifacts. Retaining these values in the figure could easily mislead readers.

(Remarks on code availability)

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Referees' comments:

Referee #1 (Remarks to the Author):

General comments

The authors address a key issue surrounding the use of point-source satellite methane detection and quantification – the lack of large-scale independent, controlled release testing. The authors conducted a year-long single-blind release study testing both different satellite sensors and different retrieval algorithms from different participants. The paper examines detection and quantification performance and discusses reasons for differing performance across 3 broad types of satellite instrument.

This is an important topic with broad reach beyond the satellite retrieval community. There is a lot of public interest in the use of point-source satellite data and I have encountered two extremes outside of academia: those who see it as a panacea for all evils and those who dismiss it as highly inaccurate and functionally useless. Fortunately, this work presents an objective evidence base to inform a more nuanced interpretation of the merits and limitations.

The scale of this work should not be underestimated. It is a highly impressive accomplishment to manage so many releases across a full year and to coordinate across the overpass times of numerous satellites and the number of participating quantification teams.

Response: We thank the reviewer for these encouraging and helpful comments!

By and large the methods and paper structure follows a well-established protocol from the same group outlining the release approach, an evaluation of detects and non-detects and a quantification parity plot for each satellite. The paper concludes that satellite measurements can allow accurate detection and unbiased quantification.

However, there are a long list of caveats to that statement, and if there's one criticism it's that these are not more front and center throughout. Given the significant investment and interest in satellite plume detectors, it is easy to see how the concluding statements may start to runaway from the realms of scientific nuance like an unstoppable freight-train on the captive eyes of the unsuspecting stakeholder.

Examples of these caveats include the still relatively small sample sizes for individual satellites (despite the impressive scale of the experiment), sanitized testing conditions and teams able to run at higher sensitivity modes than they normally would. These are discussed by the authors, but don't make it into the abstract or some of the more sweeping statements within the paper. The other main point is to emphasize in the broad concluding statements is that these findings are

specific to plume detectors and the findings relate only to the detection of large individual plumes, not a finding on the ability of satellites to accurately quantify emissions more broadly.

Response: We appreciate and agree with the reviewer's suggestions to properly frame this study's caveats and applicability. We've made several changes to the abstract and to the concluding statements throughout the paper.

Updated abstract text: "Our findings show that satellite-based measurements can achieve accurate detection and quantification of methane point-sources when the source location is approximately known."

Updated final paragraph text: "We have shown that purpose-built sensors can robustly detect methane emissions near 50-200 kg/h in ideal conditions, while numerous teams can also wield repurposed instruments to detect and quantify methane with significant skill."

Updated discussion text: "Ideally, the instrument- and team-specific quantification fits developed here would be useful for characterizing each method's performance and adjusting estimates of real-world emissions. However, the applicability of test-based calibration factors to globally measured plumes requires further study. Future validation campaigns should be prioritized to include: 1) locations with varying surface terrains and albedos; 2) locations with different weather conditions and sun availability; and 3) offshore releases to test performance over water."

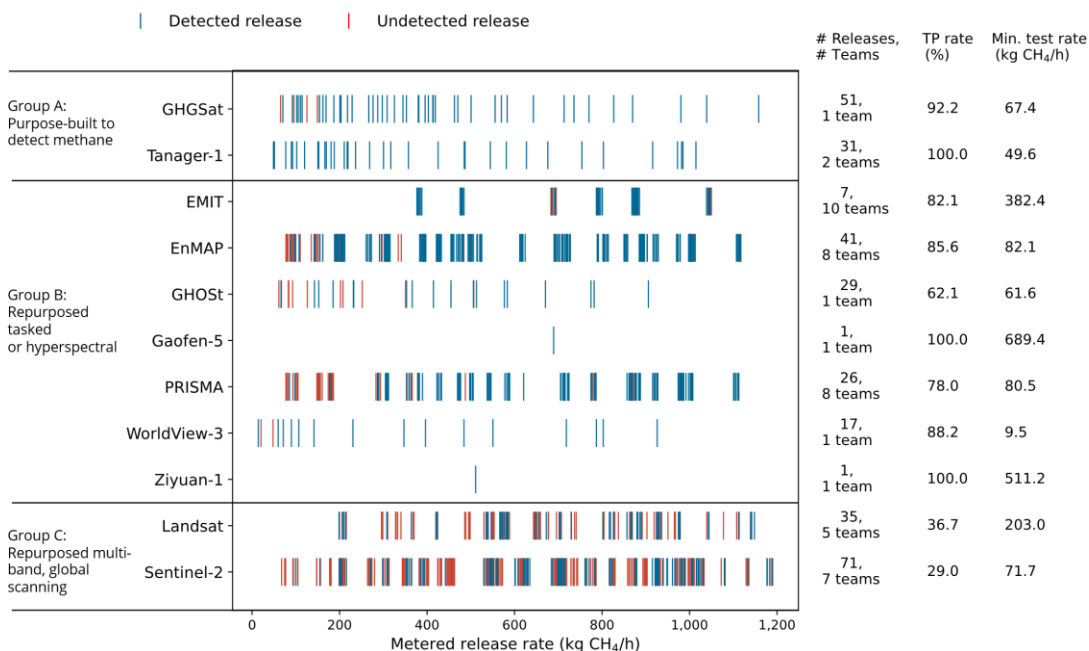
The experiment design used here, where participants know the approximate source location, likely best represents a targeted monitoring context where analysts are examining a known area of interest. However, many plumes are discovered through passive, wide-area scanning with no a priori information. Therefore, a "fully blind" test that is completely unknown to participants would be the best validation of these survey-level workflows.

While critical, controlled release testing has practical limitations, including environmental impacts and satellite revisit times that constrain sample sizes. Therefore, focused validation efforts should be complemented by large-scale statistical analyses of operational data acquired from global surveys. Such "in-the-wild" datasets lack ground truth but offer vast sample sizes to drive robust performance statistics across a global range of conditions. "

Some presentation of results may be inadvertently open to misinterpretation – for instance a 100% TP rate within a 0-250kg/h bracket may give the false impression that a particular offering can cover this full range with 100% accuracy. The reality is that the sample size is small, most of the testing in this bracket appears to have been conducted above 100 kg/h and the smallest non-zero release may have been around 50 kg/h. The danger is that the reader is unaware of the fine print and walks away with an overly optimistic outlook.

Response: We have added a column to Figure 1 with the smallest non-zero rate tested by satellite. And we have updated the text to bring the TP rates into context with the lowest tested rates.

Updated figure:



Updated text: “These results suggest strong evidence for reliable Y/N detection of sources ~50-200 kg CH₄/h by Group A and some Group B satellites in idealized conditions.”

Perhaps the most useful insights are the Group C as they have relatively good coverage across the year and the higher detection limits ensure a far greater proportion of testing occurred within the a priori marginal detection range. But curiously, the probability of detection is not investigated in any detail.

Response: We thank the reviewer for their point regarding the Probability of Detection (POD) curves. To provide insight into detection performance, we have created Figures S2 and S3, which detail detection success by release rate, aggregated by satellite and team. We did not attempt to fit formal POD curves for several reasons (1) methodological heterogeneity: given the significant differences in methods and results between teams, aggregating their results into a single POD curve for a given satellite would be inappropriate and misleading; (2) insufficient sample size: for individual teams, the sample sizes are not large enough to support reliable statistical curve fitting—an extreme case is Tanager-1, whose 100% TP rate in this study makes fitting a meaningful POD curve impossible; (3) group C instruments were tested primarily <1000 kg/h which is in their

partial detection range. This would need to be combined with outside data across a broader emissions range to produce a fuller POD picture.

While fitting POD curves was beyond the scope here, we provide data that can contribute to more sophisticated external analyses of POD. We are pleased to note that our data has already contributed to such work. For example, a recent preprint by Ramier et al. leverages data from multi-campaign controlled releases, including this one, to build a probabilistic, conditions-dependent detection model for GHGSat. We have updated the manuscript text to explain this and have added a reference to this study, directly linking our work to the community's progress on this front.

Updated main text: “Recent analyses find that GHGSat’s 50% POD (Q_{50}) from multi-campaign controlled releases (including this study) is near 100 kg CH₄/h, a factor of ~4.5 lower than detections from global surveys (Q_{50} = 460 kg CH₄/h (ref 1))”

Updated SI text: 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.

- 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 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.*

However, this dataset is a valuable input for external studies focused on POD modeling. For instance, our data contributed to a 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 assessments.

- (1) Ramier, A. et al. Performance validation of the GHGSat Methane Constellation via controlled releases. Preprint at <https://doi.org/10.31223/X5N78R> (2026).*

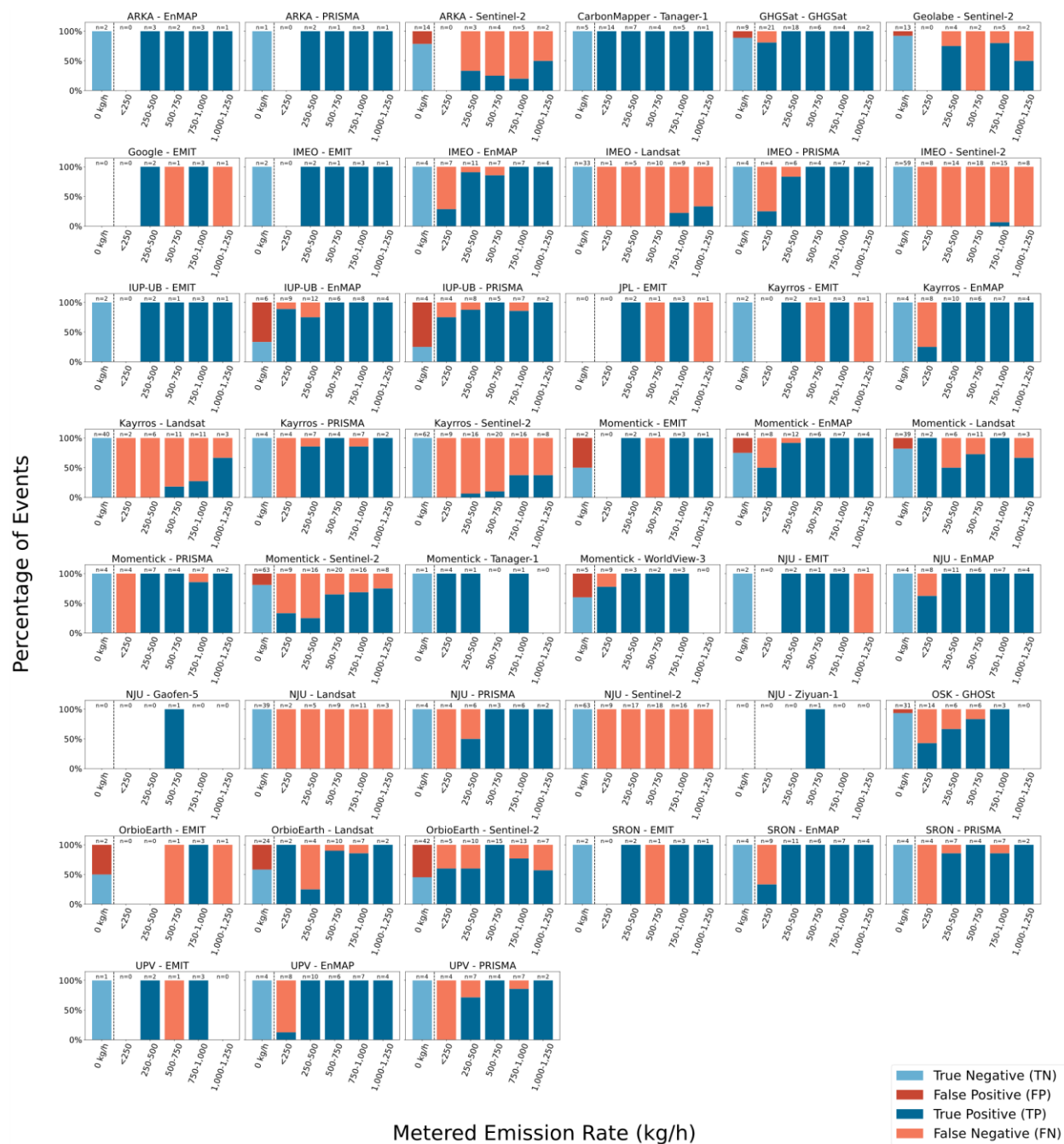


Figure S2. Detection performance binned by release rate. Bars left of the dashed line correspond to zero release events (blanks), while bars to the right correspond to non-zero releases. Some teams participated in different study phases and filtered different releases, leading to different counts for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

The lack of involvement of the participating quantification teams in the final analysis is a missed opportunity. There is no sense of scientific improvement or endeavor towards some greater sense

of verisimilitude. It's clear that subjective decisions can greatly influence estimated emissions from the open data satellites such as EMIT and Sentinel-2. Two teams presented with the same data can arrive at very different results. But the paper never really gets at why. Can this paper point towards ways to reduce the reliance on subjectivity?

Response: The reviewer raises an excellent point regarding the underlying causes of variance in satellite-based methane quantification. We agree that investigating the role of subjective investigator decisions is important and is a major, recognized challenge. Cross-team differences are influenced by many team-specific methodological choices including retrieval algorithms, signal-to-noise thresholds, plume masking approaches, and the choice of wind reanalysis product. To address this, the remote sensing community has initiated large-scale efforts like the ESA Project MEDUSA (1) and the Data Foundations Technical Working Group (2) specifically to decompose these sources of uncertainty.

A fundamental barrier for these large-scale efforts is the lack of well-controlled datasets suitable for decomposition. Our study helps fill this gap, and the dataset is already being used by the community to disentangle the factors the reviewer highlights.

A full investigation into the "why" is beyond the scope of this paper because many methods from commercial teams are proprietary, which limits immediate explanatory power. Our study design respected this reality to ensure broad participation. Furthermore, it was important for the experiment team/author group to maintain complete independence in analysis of results. This is a precedent set by past controlled release studies from multiple research groups (3, 4).

We've added a section to the SI with some explanation. And in the main text, we've explicitly pointed readers to key publications and intercomparison analyses:

(1) 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->.

(2) Akani-Guéry, J., Roger, J. & Buchwitz, M. ESA Project MEDUSA: Algorithm Intercomparison Report (AIR). (2026).

Other refs:

(3) Sherwin, E. D. *et al.* Single-blind test of nine methane-sensing satellite systems from three continents. *Atmospheric Meas. Tech.* 17, 765–782 (2024).

(4) 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).

Updated main text: “Inter-team variance in quantification is an active research area, with multiple intercomparison efforts underway outside of this study to examine the drivers of high-variance cases (ref 1, ref 2)”

Updated SI text: “It is well-documented that different analysis teams can produce divergent quantification results from identical satellite data. This variance stems from diverse methodological choices, including retrieval algorithms, data filtering thresholds, plume masking, and wind data selection. Recognizing this challenge, the remote sensing community has launched major intercomparison initiatives, such as the ESA Project MEDUSA and the Data Foundation’s Technical Working Groups, which aim to systematically decompose these sources of uncertainty.

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 study contributes to these community 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.”

Updated SI text: “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 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 (ref 3, ref 4).

None of this should take away from the importance and quality of the scientific work. Merely the implications of its publication perhaps warrant a little more oversight of its messaging.

Specific comments

Abstract – Shouldn’t the abstract mention the range across which the detection statistics apply? Because they are specific to the distribution of release magnitudes of this test.

Response: We added the release magnitude range to the summary paragraph.

Updated text: “We conducted a year-long, single-blind test involving 746 events (481 methane releases, 9.5 to 1183.7 kilograms per hour, and 265 blank controls).”

Abstract L17 – What does 15 times larger mean in this context? Release size, number of releases or number of participants?

Response: We removed the “15 times” and relocate the “largest” to refer to number of analysis teams, breadth of international participation, and estimates generated.

Updated text: “Fifteen analysis teams from nine countries generated >2,250 estimates, making this the largest controlled test to date and the first evaluation of EMIT, Tanager-1 (Carbon Mapper), and GHOSSt.”

Abstract L27-28: “Broad land scanners performed poorly but have consistent global coverage and vast historical archives, motivating additional work”. This needs further clarification – their performance is worse for detection undoubtedly. But for those plumes that they do detect, the quantification accuracy doesn’t appear to drop off markedly compared to Group A and Group B.

Response: We updated the sentence to clarify how poor Group C performance is specific to detection at this study’s tested rates. Due to space limitations in the abstract, we do not discuss quantification.

Updated text: “Broad land scanners detect poorly at tested rates but have consistent global coverage and vast historical archives, motivating additional work.”

P2. L67 – Isn’t IMEO part of UNEP rather than the UN?

Response: We’ve changed it to UNEP.

Updated text: Thousands of detected plumes are reported through public data portals, including those managed by the UNEP International Methane Emissions Observatory (IMEO)...

P5. L140 – Is IMEO a research institution?

Response: IMEO operates similarly to a research group by conducting scientific studies and managing public data platforms, but we agree they are best categorized as an intergovernmental organization, which we’ve added.

Updated text: “Analysis teams included nonprofit Carbon Mapper; intergovernmental organization the International Methane Emissions Observatory (IMEO)...”

P5. L154-156 – Is there any significance to the phasing of the experiments across the year? Did any team update their algorithm based on the ground truth data of a previous phase and improve in subsequent phases?

Response: We chose a quarterly phase structure to provide somewhat fast data turnaround for analysis teams while keeping a reasonable workload for the experiment team. It

was also intended to enable methodological updates by analysis teams. However, most teams did not report any methodology changes throughout the experiment. We've added two sentences to Experimental Design section to highlight this point.

Updated text: "While the quarterly data unblinding may have enabled teams to iterate on their methodology, most did not report any methodological changes throughout the experiment. Any reported changes are discussed in SI S2.1 and S2.2"

P10. L227 – "These FPS can be more or less impactful". Suggest this point could be more forceful. FPs are the fastest way to burn credibility.

Response: We strengthened the wording of this point.

Updated text: "Thus, aggressive algorithm tuning likely allows detection of smaller emissions sources (higher TPs) but can cause artifacts to be detected as plumes (higher FPs). This trade-off should be considered carefully as FPs can severely undermine measurement credibility and lead to wasted repair efforts."

P11. L245 – Does this have to be sidelined into future work? Even for this one example, wouldn't this paper offer so much more if it investigated why the same satellite retrieval can result in such different solutions? Investigator decisions are clearly important with these indirect forms of measurement. How can this work point to ways to reduce the impacts of subjective investigator decisions?

Response: The reviewer raises an excellent point regarding the underlying causes of variance in satellite-based methane quantification. We agree completely that investigating the role of investigator decisions is important and is a recognized challenge. Cross-team differences are influenced by many team-specific methodological choices, including retrieval algorithms, signal-to-noise thresholds, plume masking approaches, the choice of wind reanalysis product, and the modulation of wind speed. To address this, the remote sensing community has initiated large-scale efforts like the ESA Project MEDUSA and the Data Foundations Technical Working Group specifically to decompose these sources of uncertainty. A full investigation into the "why" is beyond the scope of this paper because many methods from commercial teams are proprietary, which limits immediate explanatory power. Our study design respected this reality to ensure broad participation.

Regarding how our work can support studying these questions: a fundamental barrier for these large-scale efforts is the lack of well-controlled datasets suitable for decomposition. Our study helps fill this gap, and the dataset is already being used by the community to disentangle the factors the reviewer highlights.

We've added a section to the SI with some explanation. And we conducted a simple ensemble analysis that combines outputs across multiple teams. And in the main text, we've explicitly pointed readers to key publications and intercomparison analyses:

(1) 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->.

(2) Akani-Guéry, J., Roger, J. & Buchwitz, M. ESA Project MEDUSA: Algorithm Intercomparison Report (AIR). (2026).

Updated SI text: "It is well-documented that different analysis teams can produce divergent quantification results from identical satellite data. This variance stems from diverse methodological choices, including retrieval algorithms, data filtering thresholds, plume masking, and wind data selection. Recognizing this challenge, the remote sensing community has launched major intercomparison initiatives, such as the ESA Project MEDUSA and the Data Foundation's Technical Working Groups, which aim to systematically decompose these sources of uncertainty.

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 study contributes to these community 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."

Updated main text: "Inter-team variance in quantification is an active research area, with multiple intercomparison efforts underway to examine the drivers of high-variance cases (ref 1, ref 2). One strategy for reducing uncertainty may be meta-analyses that combine outputs from multiple teams. Our ensemble analysis shows that aggregating independent estimates can significantly improve quantification precision (see SI S3.9.3)."

Updated SI S3.9.3: "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).

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.

We then calculated the signed percent error (to assess systematic 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.

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.

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."

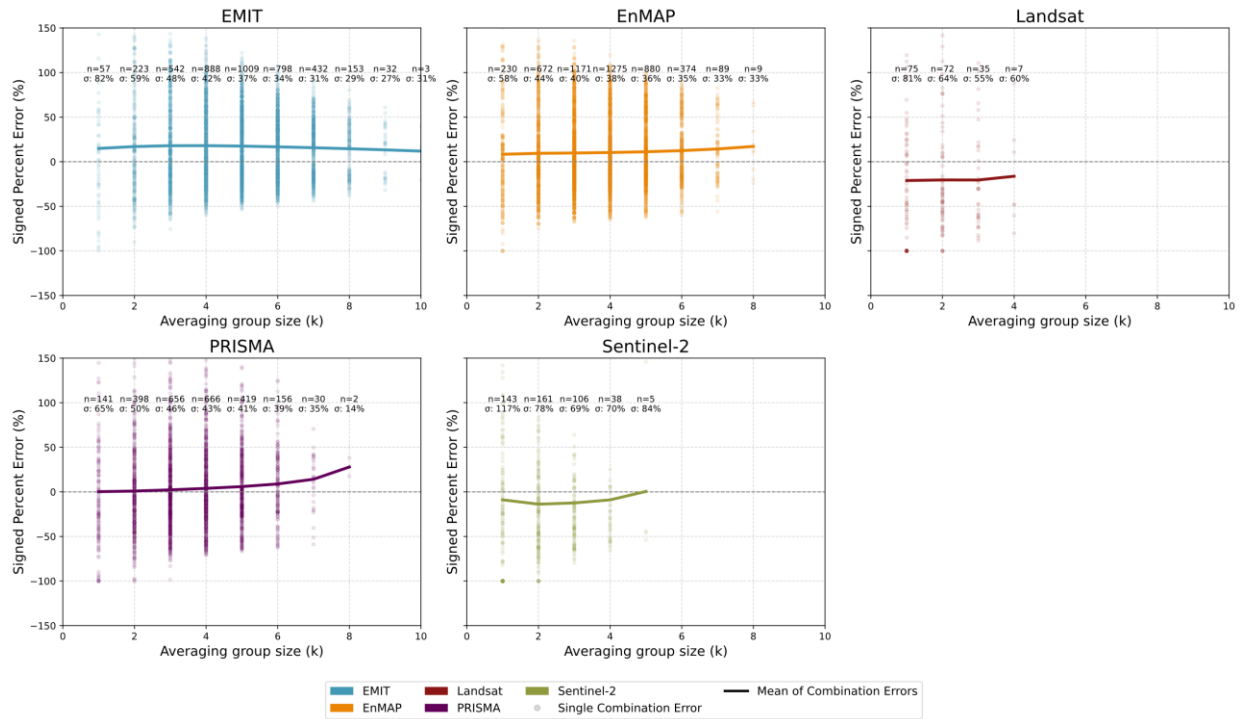


Figure S80. Signed percent error across averaging group sizes k . For each release event, we evaluated the effect of creating an ensemble of k teams, where k ranged from 1 to the maximum number of participating teams (N). For each k , all possible combinations of teams were generated, and the mean of 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.

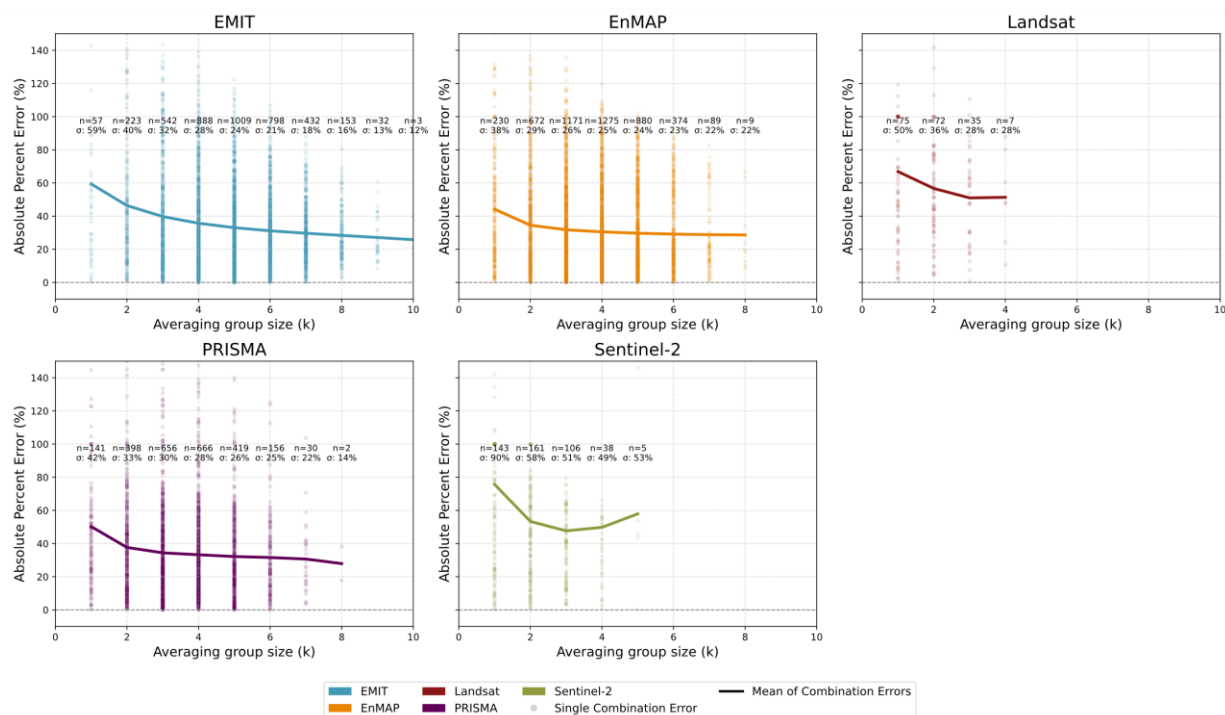


Figure S81. Mean absolute percent error across averaging group sizes k . For each release event, we evaluated the effect of creating an ensemble of k teams, where k ranged from 1 to the maximum number of participating teams (N). For each k , all possible combinations of teams were generated, and the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

P17. L375 - This bin of 200 kg/h is too open ended and gives the wrong impression. Most of these satellites were not tested below 50 kg/h and there is clearly a lower bound above 0 on which this statement might apply.

Response: We've updated the lower bound to 50 kg/h given that was the lower range of emissions tested in this study.

Updated text: "These results suggest strong evidence for reliable Y/N detection of sources ~50-200 kg CH₄/h by Group A and some Group B satellites in idealized conditions"

P17. L 377 – Is there a citation for this GHGSat Permian evaluation? Not clear how this 50% POD detection limit was arrived at.

Response: The original reference was by Worden et al. (1), but we've updated to more recent publications by Jervis et al. (2) and Ramier et al (3)

(1) Jervis, D., MacLean, J.-P. W., McKeever, J., Ramier, A. & Strupler, M. Joint Estimation of the Global Facility-Scale Oil–Gas Methane Emission Rate Distribution and Survey-

Mode Detection Probabilities from Multiple Satellite Systems. Environ. Sci. Technol. 60, 14961–14970 (2026).

(2) Ramier, A. et al. Performance validation of the GHGSat Methane Constellation via controlled releases. Preprint at <https://doi.org/10.31223/X5N78R> (2026)

(3) Worden, J. R., Nickeson, J. & Policelli, F. CSDA GHGSat Evaluation PI Report v2. Commercial Satellite Data Acquisition Program GHGSat Principal Investigator Evaluation Summary <https://earthdata.nasa.gov/s3fs-public/2025-03/CSDA%20GHGSat%20Evaluation%20PI%20Report%20v2.pdf> (2024)

Updated text: “But test performance is unlikely to be totally representative of performance in the field. Recent analyses find that GHGSat’s 50% POD (Q_{50}) from multi-campaign controlled releases (including this study) is near 100 kg CH_4/h , a factor of ~ 4.5 lower than detections from global surveys ($Q_{50} = 460 \text{ kg } \text{CH}_4/\text{h}$) (ref 2, ref 3)”

P17. L375-389 – This is a long list of important caveats. And ultimately it would be very hard to account for every one through controlled release testing. Does this point to natural limits on the use of release testing? Potentially statistics such as probability of detections may need to be derived via other means where the ground-truth is not necessarily controlled but larger sample sizes can be collected. This is somewhat implied in the discussion of Tanager detection distribution in Section S3.6.

Response: Yes, we agree. We’ve added a paragraph in the discussion to point out where large scale statistical analyses can complement CR tests.

Updated text: “Ideally, the instrument- and team-specific quantification fits developed here would be useful for characterizing each method’s performance and for adjusting estimates of real-world emission sources. However, the applicability of test-based calibration factors to globally measured plumes requires further study. Future validation campaigns should be prioritized to include: 1) locations with varying surface terrains and albedos; 2) locations with different weather conditions and sun availability; and 3) offshore releases to test performance over water.

The experiment design used here, where participants know the approximate source location, best represents a targeted monitoring context where analysts are examining a known area of interest. However, many plumes are discovered through passive, wide-area scanning where no a priori information exists. Therefore, a “fully blind” test that is completely unknown to participants would be the best validation of these survey-level workflows.

While critical, controlled release testing has inherent practical limitations, including environmental impacts and satellite revisit times that constrain sample sizes. Therefore, these focused validation efforts should be complemented by large-scale statistical analyses of

operational data acquired from global surveys. While such “in-the-wild” datasets lack ground truth, they offer vast sample sizes to drive robust performance statistics across a global range of conditions.”

P17. L392-393 – The theoretical detection limit is estimated at 1000 kg/h for S2. Perhaps it needs mentioning that 303 of 345 detection outcomes (88%) took place below this simulated detection limit. In other words, the poorer detection performance is entirely expected based on a priori knowledge. I assume this is a much higher percentage than Groups A and B. Does this affect the interpretation of statistics across groups? For comparison what percentage of testing took place below the simulated (or empirically derived) detection limits of the other satellite groups? Assuming this is small, is this why the paper does not explore probability of detection curves? For plume detecting satellites to be used for quantification purposes across broader timescales don't we need more empirical data on this?

Response: We thank the reviewer for their point regarding the Probability of Detection (POD) curves. The derivation of POD is a complex undertaking, and developing a robust methodology for this analysis was outside the paper scope.

To provide insight into detection performance, we have created Figures S2 and S3, which detail detection success by release rate, aggregated by satellite and team. We did not attempt to fit formal POD curves for several reasons (1) methodological heterogeneity: given the significant differences in methods between teams, aggregating their results into a single POD curve for a given satellite would be inappropriate and misleading; (2) insufficient sample size: for individual teams, the sample sizes are not large enough to support reliable statistical curve fitting—an extreme case is Tanager-1, whose 100% detection rate in this study makes fitting a meaningful POD curve impossible; (3) Group C instruments were tested primarily <1000 kg/h, which is in their partial detection range. These results would need to be combined with additional data to produce a fuller POD picture.

Our study provides data that can contribute to more sophisticated external analyses of POD. We are pleased to note that our data has already contributed to such work. For example, a recent preprint by Ramier et al. leverages data from multi-campaign controlled releases, including this one, to build a probabilistic, conditions-dependent detection model for GHGSat. We have updated the manuscript text to explain this and have added a reference to this study, directly linking our work to the community's progress on this front.

Updated SI text: 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.

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 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.

However, this dataset is a valuable input for external studies focused on POD modeling. For instance, our data contributed to a 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 assessments.



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.

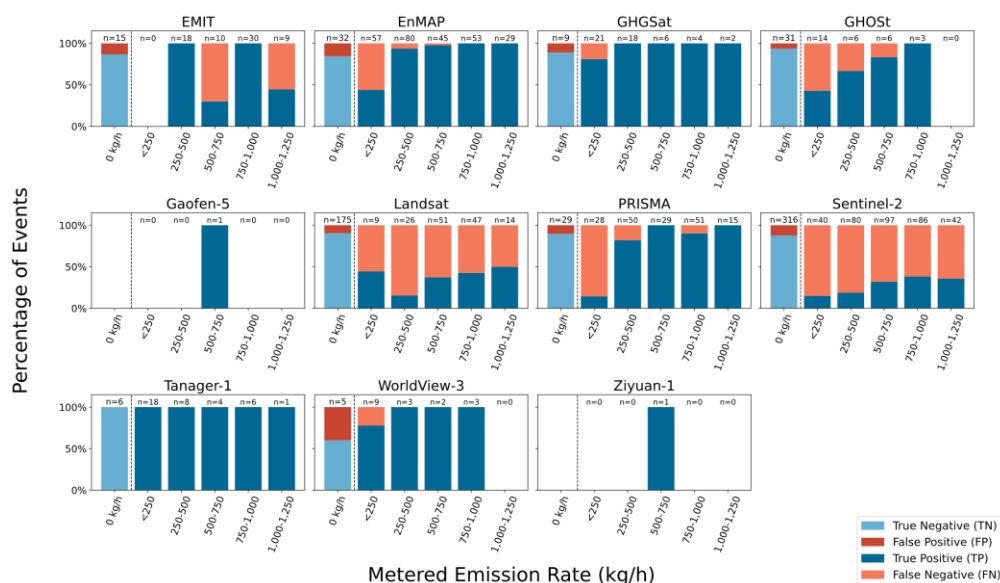


Figure S3. Detection performance aggregated by satellite, binned by release rate. Bars left of the dashed line correspond to zero release events (blanks), while bars to the right correspond to non-zero releases. Counts represent number of releases and analyzing teams (e.g. 2 releases x 9 teams analyzing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

P18. L405 – The relationship between the strictness of detection criteria and the application of the quantification team seems quite pertinent. It’s noticeable that commercial participants tend to set more lenient criteria (resulting in more FPs) than say an operational alert program. FPs are the fastest way to burn credibility and have the potential to jeopardize methane mitigation initiatives. Perhaps this point needs to be made more strongly.

Response: True. We updated the sentence to state that most FPs came from commercial teams’ methods. And we added the below table to the SI

Updated text: “Most false positives (79%) in this study originated from commercial teams applying AI-based methods to Sentinel-2 and Landsat data, leading to low precision values (see SI Tables S6, S7 and section S3.5 confirming statistical significance of higher FPs). While there is balance to be struck with avoiding missed plumes (false negatives), a system that frequently

generates false leads risks measurement credibility and is operationally untenable for many applications.”

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
Kayrros	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 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.*

P19. L415 – Any reason to pick out Nigeria and Venezuela from oil-producing tropical countries? Are they particularly plume-heavy relative to other tropical producers?

Response: Our choice was based on the intersection of high emission rates and significant monitoring challenges. According to the IEA Methane Tracker, Nigeria and Venezuela are the highest methane emitters from the oil and gas sector in their respective regions. Second, these tropical regions are known for persistent cloud cover, which presents a major obstacle for satellite monitoring. Studies have documented significant data sparseness over Venezuela for instruments like TROPOMI due to clouds.

We have revised the SI text to make this justification explicit for the reader.

References:

- (1) <https://www.iea.org/data-and-statistics/data-tools/methane-tracker>
- (2) <https://acp.copernicus.org/articles/21/14159/2021/#&gid=1&pid=1>

Updated SI text: “This 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 (ref 1, ref 2). For context, Nigeria is the largest oil producer in Africa, while Venezuela is one of the largest producers in South America.”

P18. L417 – In this experiment participants have the advantage of knowing the source location. This is akin to an operational situation where the target is known and tasked specifically. But for the other operational case of unknown unknowns would participants achieve successful detections in >45% cloud cover without this knowledge?

Response: We removed the sentence about detections made during cloud cover >45% due to have very limited data samples. We added a paragraph of discussion about cloud cover challenges to the SI.

Updated SI text: Our study provides empirical data on methane detection under cloudy conditions, a topic with limited published literature. The findings must 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 unspecified (ref 1). Disentangling the interdependent effects of cloud fraction, cloud geometry, and prior knowledge remains an area for future investigation.

(1) <https://carbonmapper.org/articles/seeing-around-the-clouds>

P19. L426-429 – The point about plume detecting satellites being important in basins where sources >100kg/hr dominate is fine, but is the Permian a good example? For instance, Williams et al (2025) arrived at an opposite distribution where 70% emissions are below 100 kg/h rather than above. Appreciate this is an area of debate but given that, it might be more appropriate to keep the point abstract rather than specific.

References

Williams et al. (2025) Small emission sources in aggregate disproportionately account for a large majority of total methane emissions from the US oil and gas sector, Atmos. Chem. Phys., 25, 1513–1532, <https://doi.org/10.5194/acp-25-1513-2025>

Response: We have abstracted this point to a high-emitting basin where large sources dominate.

Updated text: “A sensor (or constellation of sensors) with a 100 kg CH₄/h detection limit could capture a large fraction of emissions in a high-emitting basin where large sources dominate if coverage was sufficiently high. It would detect a much smaller emissions fraction in a basin with numerous, smaller sources (ref 1)”

(1) Sherwin, E. D. et al. US oil and gas system emissions from nearly one million aerial site measurements. Nature 627, 328–334 (2024).

P19. L434-436 – Does it really matter whether a controlled release targeting satellites can cover this, that and every other detail? There will always be aspects of real life that can’t be controlled. And should we try? The scale of this experiment is mightily impressive and yet the returned sample sizes remain relatively small for each individual satellite because of revisit times and quality control. As mentioned above, perhaps the solution is to sacrifice near-perfect emissions knowledge for larger statistical sample sizes.

Response: Yes, we agree. We’ve added a paragraph in the discussion to point out where large scale statistical analyses can complement CR tests.

Updated text: “Ideally, the instrument- and team-specific quantification fits developed here would be useful for characterizing each method’s performance and for adjusting estimates of real-world emission sources. However, the applicability of test-based calibration factors to globally measured plumes requires further study. Future validation campaigns should be prioritized to include: 1) locations with varying surface terrains and albedos; 2) locations with different weather conditions and sun availability; and 3) offshore releases to test performance over water.

The experiment design used here, where participants know the approximate source location, best represents a targeted monitoring context where analysts are examining a known area of interest. However, many plumes are discovered through passive, wide-area scanning where no a priori information exists. Therefore, a “fully blind” test that is completely unknown to participants would be the best validation of these survey-level workflows.

While critical, controlled release testing has inherent practical limitations, including environmental impacts and satellite revisit times that constrain sample sizes. Therefore, these focused validation efforts should be complemented by large-scale statistical analyses of operational data acquired from global surveys. While such “in-the-wild” datasets lack ground truth, they offer vast sample sizes to drive robust performance statistics across a global range of conditions.”

P19. L441 – It is hard to agree with the characterization of the emission rates tested by Group A in this study as “relatively small”. They could easily be called “exceedingly large”. The majority

are above 100 kg/h, commonly used as the arbitrary super-emitter definition. And as Fig S4 shows, the real-world detection performance certainly drops off below 100 kg/h.

Response: We changed the sentence to avoid arbitrary small/large delineations.

Updated text: “We have shown that purpose-built sensors can robustly detect methane emissions near 50-200 kg/h in ideal conditions, while numerous teams can also wield repurposed instruments to detect and quantify methane with significant skill.”

Supplement Table S3 – A number of teams used HRRR wind data – a product only available for CONUS. What would those teams use in other parts of the world and is there any indication of how this model compares to global reanalysis products such as ERA5 and thus whether similar performance would be expected for equivalent arid sites in other parts of the world?

Response: Teams that use HRRR for U.S.-based analyses would typically use a high-quality global reanalysis product such as ERA5. Regarding comparison between these models, Millstein et al. (2023) show that biases between HRRR and ERA5 can differ by region and time of day. Their analysis of the arid ERCOT region in Texas—a reasonable proxy for our study site—showed good agreement between the models during relevant midday overpass times. More directly, a recent study by Ramier et al. (2026) used both HRRR and ERA5 for GHGSat emission estimates in North America and found that the two wind sources yielded broadly comparable performance. While we cannot guarantee identical performance in all arid regions globally, it is reasonable to expect that teams using a global model like ERA5 would achieve similar results.

We have updated the SI text to include this discussion and the relevant citations.

Updated SI text: The use of the HRRR model, which is limited to the Continental U.S. (CONUS), 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) (1). Ramier et al. used both HRRR and ERA5 for GHGSat emission estimates in North America also found broadly comparable performance (2).

(1) 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).

(2) Ramier, A. et al. Performance validation of the GHGSat Methane Constellation via controlled releases. Preprint at <https://doi.org/10.31223/X5N78R> (2026).

Table S3 - ERDDAP is a data server containing multiple datasets rather than a reanalysis product – what actual dataset was used?

Response: The ARKA team reported obtaining wind data from ERDDAP. They did not specify the data product used and did not respond to follow up emails. We added a note about this in Table S3. As seen in Figure S7, ARKA’s submitted windspeeds do not line up 1:1 with other teams’ wind data, thus we treat it as a unique wind data source.

Updated SI text: “ERDDAP is a data server hosted by NOAA. The reanalysis product wasn’t specified by ARKA.”

Table S3 – Is OpenWeatherMap a unique reanalysis product or a commercial data server hosting specific reanalysis datasets?

Response: OpenWeatherMap aggregates data from multiple sources, including weather stations, ERA5 reanalysis product, and its own proprietary modeling. We added a note and a citation about this in Table S3. As seen in Figure S7, GHGSat’s submitted windspeeds do not line up 1:1 with other teams’ wind data, thus we treat it as a unique wind data source.

Updated SI text: “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.”

Referee #2 (Remarks to the Author):

The authors present the results of a year-long single-blind controlled release experiment evaluating satellite observations of methane point source plumes. The experiment evaluates 11 satellite systems, with data for hundreds of controlled releases (and zero-emission blanks) analyzed by 15 satellite teams from a variety of institutions and countries. This is by far the largest-scale controlled release trial for satellites to-date and a very important study for the methane remote sensing and mitigation communities. The key finding is that current satellite technologies work, but with a hierarchy of performance. Purpose-built methane sensors show excellent detection/quantification capabilities, while the performance of more general/repurposed instruments is at best comparable but overall moderately degraded, and that of global land imagers with very coarse spectral resolution but wide spatiotemporal coverage is more severely degraded. The experimental design, data analysis methodologies, presentation of results, and writing are all of excellent quality and I support publication of the study in Nature.

Response: Thanks to the reviewer for these encouraging comments!

I have one “major” comment and several minor ones that I would like the authors to address:

My major comment is a suggestion. I would strongly encourage the authors to perform one additional experiment, which is to evaluate the performance of the data providers as a “crowd” for plumes analyzed by multiple teams. The paper does a very good job describing individual team/sensor performance and summarizing overall capabilities of the data provider community. But is there a “wisdom of crowds” effect to document here? In other words, when analyzing the same plume, does the ensemble estimate (mean, median, something else) from a collection of data providers outperform that from any one data provider? This would be very good to know, as it could temper some of the concerns around inter-provider variability that have been recently highlighted by the methane remote sensing user/provider communities (and highlighted in the discussion of Figure 3, where the provider mean emission rate estimate is 628 kg/h – 2x closer to the truth than the individual team estimates on average).

Response: We thank the reviewer for this insightful suggestion, which prompted us to perform a new analysis for the manuscript.

We have added a new section and two new figures (S79, S80) to the SI that explores how aggregating estimates from multiple teams for the same release event impacts overall accuracy and precision. We performed this analysis for the public satellites (EMIT, EnMAP, PRISMA, Sentinel-2, and Landsat) where we had a sufficient number of participating teams (4-10). The methodology involves creating ensembles of increasing group size (k) and plotting the resulting signed percent error and mean absolute percent error to assess bias and precision) as a function of k.

Updated text: “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).

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.

We then calculated the signed percent error (to assess systematic 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.

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.

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.”

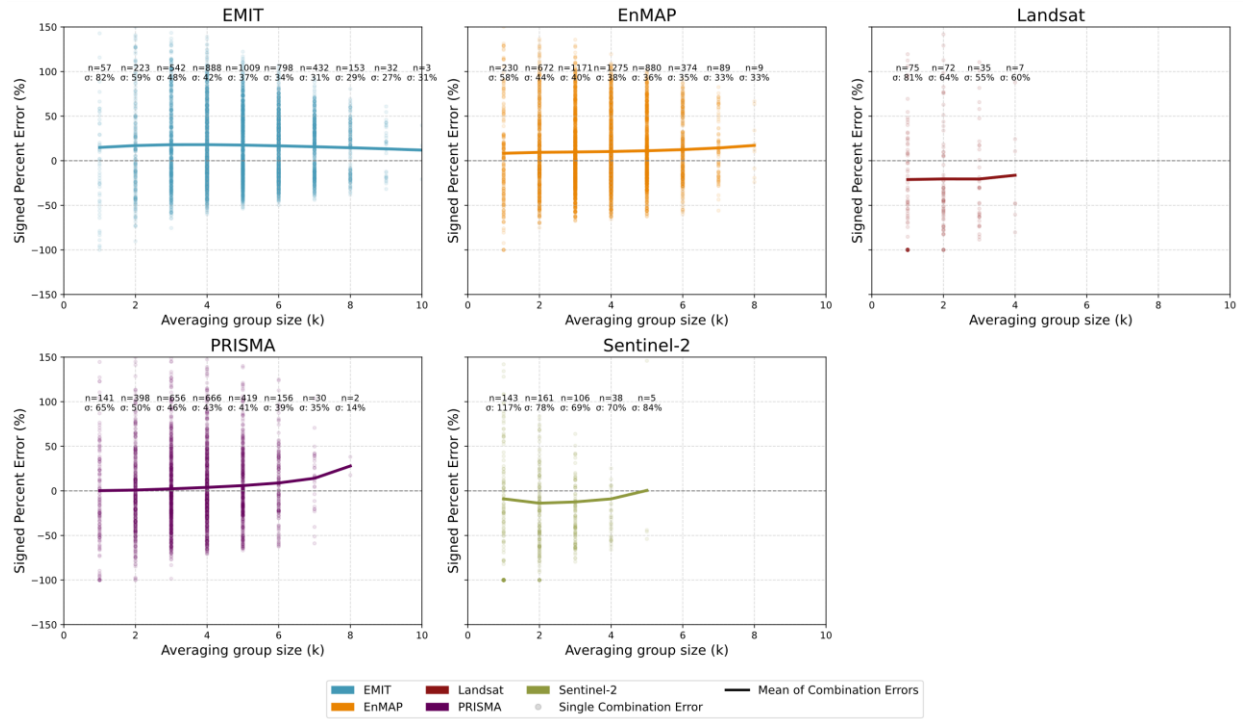


Figure S80. Signed percent error across averaging group sizes k . For each release event, we evaluated the effect of creating an ensemble of k teams, where k ranged from 1 to the maximum number of participating teams (N). For each k , all possible combinations of teams were generated, and the mean of 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.

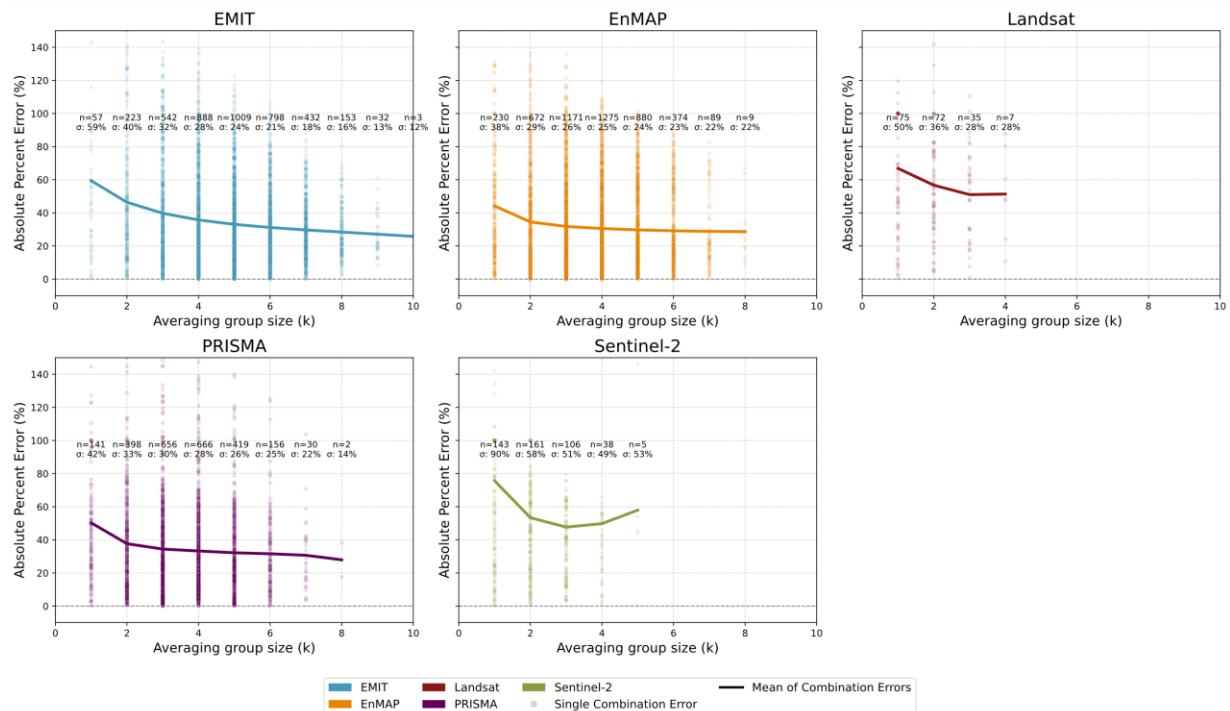


Figure S81. Mean absolute percent error across averaging group sizes k . For each release event, we evaluated the effect of creating an ensemble of k teams, where k ranged from 1 to the maximum number of participating teams (N). For each k , all possible combinations of teams were generated, and the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

Minor comments:

Line 21: "Across all sensors and teams, 70.6% correctly identified the presence or absence of emissions." I initially read the percentage as referring to the sensors and teams, but that's obviously wrong – it refers to the >2250 estimates. Rephrase for clarity?

Response: This sentence was removed in the re-written version of the summary paragraph.

Line 27: "Broad land scanners performed poorly but have consistent global coverage and vast historical archives, motivating additional work." Yes, but as mentioned several times in the manuscript, this is to be expected given the relatively low release rates (compared to land scanner detection limits). I agree wholeheartedly with the statement that follows about the motivation to improve land scanner retrievals, but also wonder whether the detection limit caveat can be highlighted here.

Response: We added in the caveat that the performance is linked to the size of the study rates.

Updated text: "Broad land scanners detect poorly at tested rates but have consistent global coverage and vast historical archives, motivating additional work."

Line 31: "Our findings show that high-quality methane measurements from space can allow accurate detection and unbiased quantification, initiating a new era in international greenhouse gas mitigation." I believe this is an overstatement and should be dialed back / rephrased. This is a very important study that can serve as a key reference for the reliability of satellite methane point source monitoring, but it does not "initiate a new era" in GHG mitigation. Satellite-based mitigation has been ongoing for quite some time already.

Response: We majorly revised the last statement of the summary paragraph per the reviewer's suggestion.

Updated text: "Our findings show that methane point-source measurements from space can allow accurate detection and quantification under favorable conditions and approximate source-

location knowledge.”

Line 47: “These instruments range from multispectral government satellites with broad Earth-observing missions (e.g. Sentinel-2) to commercial constellations of multiple satellites that offer GHG monitoring as a service.” The ESA Sentinel-2 mission overview page is cited here, which is a bit odd since it says nothing of methane. Something like the Jacob 2022 review or the Varon 2021 Sentinel-2 methane paper might be a better citation.

Response: We’ve updated this citation to Jacob 2022.

Line 52: “These are the first systems that are both designed specifically to detect methane and are committed to releasing large-scale datasets into the public domain.” A minor correction, but this statement is not really accurate (as becomes clear later in the sentence). Other instruments, like TROPOMI, have been “designed specifically to detect methane” and release “large-scale datasets into the public domain.” Carbon Mapper and MethaneSAT were the first to do this for methane point sources specifically (to my knowledge) and could also be differentiated as non-profit methane satellite missions.

Response: We’ve removed the claim that these systems were the “first.” We’ve updated the text to highlight that both are public-good missions that contributed to a step-change in methane monitoring capabilities. MethaneSAT was designed primarily for larger area sources (although it can also do large point sources), and Carbon Mapper is focused on point sources.

Updated text: “In 2024 a step-change in capability occurred with the launch of two new satellites, Carbon Mapper’s Tanager-1 and Environmental Defense Fund’s MethaneSat. Both systems are public-good missions that were designed specifically to detect methane and release large-scale datasets into the public domain, building on existing monitoring capabilities such as those from high-resolution hyperspectral imagers EnMAP and PRISMA and the coarser-resolution TROPOMI instrument (ref 1). Tanager-1 is a high-resolution hyperspectral sensor designed for point-sources, which detected more than 2,400 methane plumes globally in its first seven months of operation (ref 2). Carbon Mapper is expected to scale up its constellation with at least three more instruments to achieve daily methane monitoring of high-emitting areas (ref 2). MethaneSat, in operation from March 2024-June 2025, had finer spectral resolution and coarser spatial resolution and was designed for wide-area total emissions estimates (ref 3).”

(1) Jacob, D. J. et al. *Quantifying methane emissions from the global scale down to point sources using satellite observations of atmospheric methane. Atmospheric Chem. Phys.* **22**, 9617–9646 (2022).

(2) Duren, R. et al. *The Carbon Mapper emissions monitoring system. Atmospheric Meas. Tech.* **18**, 6933–6958 (2025)

(3) MethaneSAT. The Satellite | MethaneSAT. <https://www.methanesat.org/satellite> (2025).

Line 110: “(6) a complete year of data, allowing for more variety in weather conditions and sun angles.” I was interested to see analysis of performance by sun angle / weather conditions, but aside from cloud cover, I don’t believe this ever appears in the manuscript. Perhaps I missed it, but if not then I would suggest clarifying this 6th contribution.

Response: We created dedicated sections S3.7.1 for wind and S3.7.2 for sun angle in the SI. In the latter, we added a new figure (below) highlighting detection accuracy across varying sun angles. Due to space limitations, the figures other than cloud are placed in the SI. However, we added discussion of sun angle to the main text, alongside discussion of wind impacts on detection.

Updated main text: “Wind speed and sun angle can also affect plume detection. Group A and C satellites showed weak changes in detection efficacy as a function of wind speed, while Group B showed some sensitivity to winds > 3 m/s, although error bars were fairly wide (see SI Figure S7). Modest decline in TP rate was observed across groups at higher solar zenith angles, consistent with reduced photon availability (see Figure S10). Further examination of meteorological conditions and detection capabilities is warranted for future study.”

Updated SI figure and text:

“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.

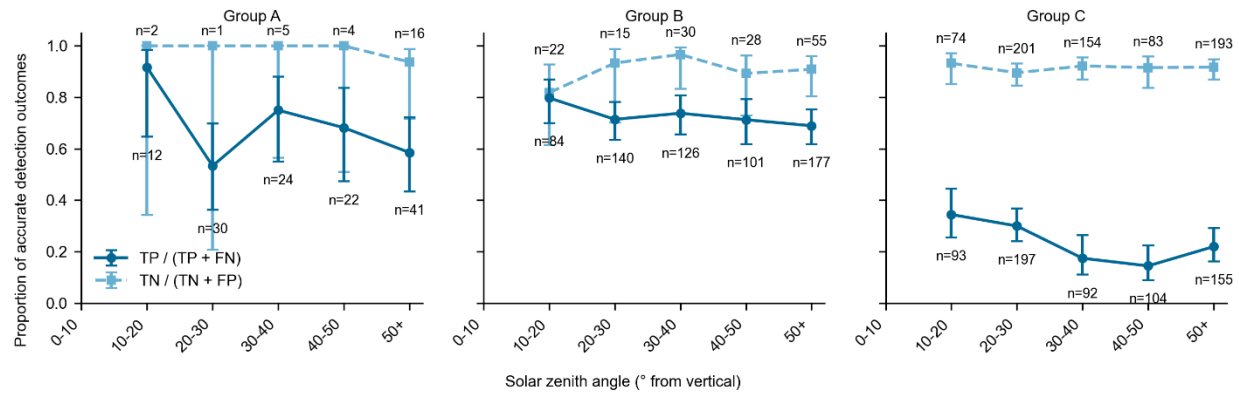


Figure S10. Proportion of accurate detection outcomes by solar zenith angle bin and instrument grouping.”

Line 131: “SRON (Space Research Organization Netherlands)”. Is this acronym intentionally reversed?

Response: We’ve corrected this.

Updated text: “...Space Research Organization Netherlands (SRON)...”

Line 169: “EMIT overpass frequency is irregular because it is mounted on the ISS and has an unpredictable non-polar orbit.” The orbit of the ISS is not unpredictable, but its precession does produce irregular revisit rates. Please rephrase for clarity.

Response: We’ve rephrased as the reviewer suggested.

Updated text: “Because EMIT is mounted on the ISS, whose non-polar orbit produces irregular revisit patterns, the number of possible controlled release observations is limited.”

Line 196: “This should not be taken as evidence of the general ability of Group C instruments to see very large (>1500 kg/h) emission sources, as our equipment can only release volumes at the expected low end of the detection range for these instruments.” Is there a good reference to make the point that the Group C instruments do routinely detect large sources globally? Perhaps something from IMEO’s MARS work could make that point.

Response: We’ve added a sentence and two references to support the point that Group C instruments effectively detect very large sources in controlled and operational settings.

Updated text: “Past controlled experiments and public data platforms have shown effective detection of very large sources (ref 1, ref 2)”

(1) Sherwin et al 2023

(2): IMEO MARS platform

Line 209: “Rates must therefore be adjusted for cloud coverage in a given location.” I do not understand this. Is it saying that the TP rate would be degraded for partially cloudy scenes? How would you adjust the rates to account for cloud cover? Please elaborate/rephrase to clarify.

Response: We’ve removed the sentence entirely to avoid confusion.

Line 217: “In contrast, outcomes for Sentinel-2 and Landsat varied substantially among teams. Out of 71 non-zero releases for Sentinel-2, two teams made 0 or 1 TP detections while two other teams made >35 TP detections (see SI Figure S2 for each team’s outcomes by release rate); Sentinel-2 F1 scores range 0.0-0.73 (SI Table S6).” I thought this was a very interesting finding and was disappointed not to get some kind of “concluding sentence” afterwards, like the preceding one stating that “the very similar TP rate” despite the “heterogeneity of analysis approaches” for Group B instruments shows that their detection limits are governed by spectral information. Does the much wider range of results for the land scanners mean that there is potentially more information available from e.g. Sentinel-2 than is accessible by standard data analysis methods, and that AI (used by some teams and not others) can unlock? Does it mean something else? What did the high-performing Group-C teams do better than the low-performing teams?

Response: We added a concluding sentence to this section.

Updated text: “...In contrast, outcomes for Sentinel-2 and Landsat varied substantially among teams. Out of 71 non-zero releases for Sentinel-2, two teams made 0 or 1 TP detections while two other teams made >35 TP detections (see SI Figure S2 for each team’s outcomes by release rate); Sentinel-2 F1 scores range 0.0-0.73 (SI Table S6). The variance suggests that for these instruments operating in their marginal signal-to-noise range, analyst choices like signal thresholding become a dominant factor.”

Line 241: “two stand out as having significantly different smoothing and brightness.” Presumably those are Google and Orbio Earth? Would be worth making explicit.

Response: We’ve updated this sentence.

Updated text: “For example, while nine of the ten teams analyzing EMIT for the 3 June case show plumes traveling downwind, the Google and Orbio Earth outputs stand out as having significantly different smoothing and brightness.”

Line 244: “There is no apparent trend between errors in assumed windspeed and enhancement concentration in this example.” It was not clear to me whether this means that the teams all used similar wind speeds to estimate the emissions. Is that the correct interpretation? Or is the sentence saying that windspeed errors and methane plume concentrations are not correlated (why would they be?)? Using different wind speeds could of course lead to large differences, as could different masking techniques. Please clarify this sentence.

Response: We’ve clarified this sentence to explain that differences in team-reported windspeeds did not explain the 4X quantification variance.

Updated text: “Even visually similar enhancements can yield substantially different source rate estimates; reported rates for this event varied by over 4x (348 kg CH₄/h [0, 1,196] to 1,406 kg CH₄/h [10, 2,803]). While windspeed is often a linear factor in rate estimation, team-reported windspeeds did not account for the spread in this example.

Inter-team variance in quantification is an active research area, with multiple intercomparison efforts underway to examine the drivers of high-variance cases (ref 1, ref 2)…”

(1) 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->.

(2) Akani-Guéry, J., Roger, J. & Buchwitz, M. ESA Project MEDSUA: Algorithm Intercomparison Report (AIR). (2026).

Figure 5: On the inclusion (or not) of FP in the emission rate regressions, I see in the supplement that both approaches were tested. But I don’t see any summary of results for the TP-only regressions. If I have not missed it, would it be feasible to include in some form in the supplement? Some of the default FP+TP regression lines going in unexpected directions are not very informative.

Response: Per the reviewer’s suggestion, we’ve added discussion of TP-only results to the SI.

Updated text: “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 R² 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.

Line 338: “Quantification was also largely unbiased. For example, Carbon Mapper successfully detected every non-filtered release (n=31) with Tanager-1 in the year-long experiment. They also reported zero false positives across five no-release controls.” The flow of ideas here, from the first to second and third sentences quoted, is not very clear to me.

Response: We shifted the “largely unbiased quantification” to early in the paragraph where Group A overall performance is summarized. Then Carbon Mapper’s performance follow the statements about detection.

Updated text: “Group A systems demonstrated the strongest detection performance and largely unbiased quantification, characterized by high TPRs and few to no false positives for releases spanning <100 kg CH₄/h to over 1,000 kg CH₄/h. Their F1 scores range 0.95 to 1, indicating ability to accurately identify both positive and negative cases (see performance metrics in SI Table S6). For example, Carbon Mapper successfully detected every non-filtered release (n=31) with Tanager-1 and reported zero false positives across five no-release controls.”

Line 378: “A recent evaluation comparing GHGSat to airborne instruments in the Permian Basin found that its 50% POD threshold is near 250 kg CH₄/h.” Please provide the citation for this evaluation. The recent Jervis 2026 paper in ES&T could be useful here – they show that field or “survey” detection probability is considerably worse than in controlled conditions: <https://pubs.acs.org/doi/abs/10.1021/acs.est.5c17089>

Response: We’ve updated this section to reference the Jervis et al. 2026 paper and the even more recent Ramier et al 2026 preprint which compares controlled performance to survey mode.

Updated text: These results suggest strong evidence for reliable Y/N detection of sources near 50-200 kg CH₄/h by Group A and some Group B satellites in idealized conditions. But test performance is unlikely to be totally representative of performance in the field. Recent analyses find that GHGSat's 50% POD (Q_{50}) from multi-campaign controlled releases (including this study) is near 100 kg CH₄/h, a factor of ~4.5 lower than a large ensemble detections from global surveys (Q_{50} = 460 kg CH₄/h) (ref 1, ref 2).

*(1) Jervis, D., MacLean, J.-P. W., McKeever, J., Ramier, A. & Strupler, M. Joint Estimation of the Global Facility-Scale Oil–Gas Methane Emission Rate Distribution and Survey-Mode Detection Probabilities from Multiple Satellite Systems. *Environ. Sci. Technol.* **60**, 14961–14970 (2026).*

(2) Ramier, A. et al. Performance validation of the GHGSat Methane Constellation via controlled releases. Preprint at <https://doi.org/10.31223/X5N78R> (2026)

Referee #3 (Remarks to the Author):

A.Summary of key results

This one-year experiment ran from 1 January 2025 to 31 December 2025, consisting of 746 observational events including 481 quantitative controlled methane release tests and 265 blank controls. Twelve on-orbit hyperspectral and multispectral satellites were arranged to observe the ground-based controlled methane release activities. Fifteen data analysis teams from nine countries produced more than 2,250 emission estimates in total. The paper analyzes and compares the true positive rates, false positive rates, biases of methane releases detection from various satellite spectral imaging sensors. It also evaluates the impacts of cloud cover on methane detection performance and compares inversion approaches driven by artificial intelligence.

B.Originality and significance

This study provides an abundant dataset of ground-based controlled methane release experiments conducted in the desert area of Casa Grande, Arizona, USA, including blank observations with zero methane emissions. The tested methane emission rates range from 9.5 to 1,183.7 kg CH₄/h, and the dataset records diverse observational conditions such as albedo, solar elevation angle, wind speed and cloud coverage. This dataset delivers substantial value to research on satellite-based methane emission detection.

Furthermore, the comprehensive detection results covering different satellites, wind speed bins, cloud coverage, emission magnitudes and independent analysis teams carry important academic and practical implications for understanding methane monitoring performance of spaceborne spectral imagers.

Response: Thanks to the reviewer for these encouraging comments!

Nevertheless, several points deserve clarification:

Four core components constitute the whole research workflow:

- ① Ground-based controlled methane release setup and concurrent meteorological measurements;
- ② Satellite overpass observations synchronized with pre-designed release/non-release test scenarios;
- ③ Over 2,250 release or non-release estimates generated from experimental data;
- ④ Integrated summary and analysis of all experimental setup and observational outcomes.

Components ② and③ represent the core contributions from satellite operators and especially the 15 independent analysis teams. Their technical and academic inputs are indispensable to the valuable dataset and meaningful detection results presented in this paper. Although key participants are acknowledged at the end of the manuscript, the current author affiliation list does

not fully capture the full scope of cross-team contributions to this large-scale collaborative field campaign.

Response: We wholeheartedly agree with the reviewer's point that the analysis teams' technical contributions are essential for this paper's results.

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 was to ensure the complete independence and scientific integrity of the final analysis. Including participating teams in the author group of the primary evaluation paper could introduce a real or perceived conflict of interest.

This model is not unique to our study; it adheres to a well-established precedent set by other major single-blind controlled release intercomparisons. This includes previous work from our own group (1) as well as studies from other leading research institutions like Colorado State University's METEC facility (2).

Furthermore, the participation agreement was designed to be mutually beneficial. While the organizing team publishes the primary, independent evaluation, all participating teams are explicitly free to use the data from their own analyses to publish their own separate papers, with their own author lists. For example, a recent preprint by Ramier et al. (3) incorporates releases from this study into a detailed probabilistic performance assessment of GHGSat. We were not involved as authors on that paper.

We've added text to the SI describing these points.

(1) Sherwin, E. D. et al. Single-blind test of nine methane-sensing satellite systems from three continents. *Atmospheric Meas. Tech.* 17, 765–782 (2024).

(2) 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).

(3) Ramier, A. et al. Performance validation of the GHGSat Methane Constellation via controlled releases. Preprint at <https://doi.org/10.31223/X5N78R> (2026).

Updated SI text: “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 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 (ref 1, ref 2).

C. Data & methodology

(1) Lines 50–55 state that analytical teams were blinded to all ground-truth data during each satellite overpass, including the occurrence of methane releases, exact release timings, emission flow rates and in-situ meteorological observations. While concealing ground-truth emission data aligns with single-blind experimental protocols, wind speed measurements should arguably have been shared with participating teams. As illustrated in Supplementary Information Figure S7 (Line 596), wind speeds retrieved independently by individual teams deviate substantially from in-situ observations. Given that wind field characteristics dominate methane plume dispersion and quantification characteristics, such large wind estimation discrepancies inevitably undermine the reliability of cross-satellite detection performance evaluations.

This study aims to quantify the native detection performance of spaceborne methane sensors, and wind speed cannot be directly retrieved by any of the satellites assessed here. Inter-team wind speed estimation errors can reach multiple-fold magnitudes, introducing unquantifiable confounding biases into sensor performance assessments and derived statistical results. Wind field retrieval represents an independent research domain requiring dedicated methodological development and validation, and should not act as an uncontrolled confounding variable in inter-satellite comparative experiments.

Response: We thank the reviewer for raising this important point about the role of wind. The decision to blind *in-situ* wind data was a deliberate element of our experimental design, and we appreciate the opportunity to clarify our rationale.

Our study was designed to assess the performance of the analysis teams' workflow, as close as possible to how they would be applied in the real world. In operational scenarios, teams rarely have access to co-located, ground-truth meteorological data. 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 design choice is supported by empirical evidence from previous studies. A recent two-stage controlled release experiment (Sherwin et al., 2024) first tested teams with blinded wind, then provided them with the *in-situ* data and had them re-process their results. The study found that unblinding the wind data led to only a modest improvement in aggregate error (from 29% to 27%). This suggests that while wind is undeniably a source of uncertainty, other factors within the analysis chain are also highly significant drivers of variability. Several large, multi-team intercomparison efforts such as the ESA's Project MEDUSA and the Data Foundation's Technical Working Groups are working to dissect these drivers (1, 2).

Given the evidence from past experiments, and that a two-stage experiment would have doubled the logistical burden, we opted for a single-stage blind design. To make this rationale transparent to the reader, we have added a new section to the Supplementary Information explicitly discussing this methodological choice and its justification.

(1) 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->.

(2) Akani-Guéry, J., Roger, J. & Buchwitz, M. ESA Project MEDUSA: Algorithm Intercomparison Report (AIR). (2026).

(3) Sherwin, E. D. *et al.* Single-blind test of nine methane-sensing satellite systems from three continents. *Atmospheric Meas. Tech.* 17, 765–782 (2024).

Updated text:

“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 error (ref 3). Given this evidence, and the substantial logistical burden of a two-stage analysis, we opted for a single-stage blind design.”

Updated section: “S2.3 Sources of methodological variance

It is well-documented that different analysis teams can produce divergent quantification results from identical satellite data. 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 MEDUSA and the Data Foundation's Technical Working Groups, which aim to systematically decompose these sources of uncertainty (ref 1, ref 2).

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 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."

(2) Table 1 contains an inaccurate parameter for the SWIR band of Gaofen-5: its spectral resolution is better than 10 nm (instead of 5 nm). The reference below can be cited for correction:

Li F, Lin K, Yan Y, et al. Comparing the performance of different hyperspectral satellite imaging spectroscopy in mapping methane point-source emissions. *Remote Sensing of Environment*, 2026, 334: 115224.

Response: Thanks to the reviewer for providing this citation. We've updated the paper accordingly.

(3) Figure 1 indicates limited observational coverage for several platforms: Gaofen-5 and Ziyuan-1 each only acquired one release event analyzed by a single team; GHGSat, GHOSat and WorldView-3 also received analysis from merely one team each; Tanager-1 data was processed by two teams only. All remaining satellites had at least five independent analysis teams involved, while MethaneSAT provided no analytical results.

Coordinating satellite overpass schedules to align with ground release events and organizing multiple teams to process datasets from diverse satellites undoubtedly poses logistical challenges. However, this is a large-scale one-year campaign with 481 valid methane release tests, intended for publication in a high-impact journal such as *Nature*. Severe under-sampling for certain satellites and highly uneven distribution of analysis teams significantly compromise data completeness and methodological robustness, two critical standards expected of landmark interdisciplinary research designed to integrate cross-institutional resources to address global research gaps.

The Gaofen-5 hyperspectral payload shares a 30 m spatial resolution with Tanager-1 yet delivers a 60 km swath width—over three times wider than Tanager-1. A complete performance characterization of Gaofen-5 would greatly advance the development and application of this class of hyperspectral methane sensors, and the absence of sufficient Gaofen-5 evaluation

constitutes a notable limitation of this otherwise valuable work.

Response: We thank the reviewer for this critical and thoughtful feedback. The reviewer correctly identifies the uneven sampling across platforms and teams as a notable limitation, and we appreciate the opportunity to provide context on the study's design, constraints, and overall scope.

The observed distribution reflects the real-world constraints inherent in a large-scale, voluntary, multi-national collaboration. We can break down the reasons for the uneven sampling into three categories:

- 1. Proprietary data access (GHGSat, GHOSat, WorldView-3, Tanager-1):** Most high-resolution commercial satellite data is behind a paywall, or in the case of Tanager-1, faces a 30-day data lag. Consequently, analysis for these platforms was primarily conducted by the satellite operators themselves or their close partners. Broader participation would have required significant data licensing fees, which was beyond the scope of this voluntary study.
- 2. Logistical & geopolitical barriers (Gaofen-5, Ziyuan-1):** We share the reviewer's view that a full characterization of platforms like Gaofen-5 would be highly valuable. We made numerous attempts to coordinate with our international counterparts to secure more observations and analyses. Unfortunately, as has been seen in past studies (e.g., Sherwin et al., 2024, which also secured only one observation each for ZY-1 and GF-5), coordinating with certain government space agencies presents persistent logistical challenges.
- 3. Voluntary participation (all teams, MethaneSAT):** The foundation of this study was the voluntary participation of the 15 analysis teams. We could not compel any team to analyze a specific dataset. MethaneSAT team lost communication with its satellite mid-study and did not submit results, a factor unfortunately beyond our control.

While we acknowledge these gaps, this campaign is, to our knowledge, the largest and most comprehensive single-blind satellite intercomparison ever conducted, by a substantial margin.

- It involved 15 independent analysis teams (compared to 6 in the next largest study).
- It evaluated 11 satellite platforms with submitted results (9 in next largest study).
- It ran for a full year, capturing diverse conditions (compared to 2 months for the Sherwin et al. study).
- It generated over 2,200 individual estimates, which is more than 15 times the number produced in any previous controlled release study.

While this study could not solve all challenges of data access and international cooperation, we believe its scale and statistical power provide immense value to the community. We've added text to the Methods describing these points.

Updated methods text: "The experiment team tracked satellite observation opportunities using a combination of email communications with participating teams, online user portals, and predictable orbital schedules. The scheduling of observations was determined by each satellite operator according to their own mission priorities and commercial considerations and was not controlled by the experiment team. For instruments requiring tasking, we generally requested as many observations as were available. Exceptions include requests for EnMAP and PRISMA, which were occasionally modulated to manage total volume, and for GHGSat, which provided a very high number of observations due to its larger constellation size.

...

Analyst participation in this study was voluntary and varied by phase. In Phase 1 ... "

(4) Figure S2 in Supplementary Information Line 466 shows that the two Tanager-1 analysis teams achieved perfect true positive detection across all release events. Similarly, certain teams working with EMIT, EnMAP and PRISMA also successfully detected every methane plume without missed signals. This creates ambiguity when interpreting overall performance statistics: it remains unclear whether cross-team disparities in detection outcomes stem primarily from analytical algorithms or intrinsic satellite sensor characteristics, which impairs the precision of inter-platform evaluation.

Figure 3 at Line 250 further demonstrates extreme variability in reconstructed methane plumes derived from ten independent teams processing identical EMIT overpass data.

The paragraph "Variable quantification performance across instrument classes" (Lines 284–291) reports that Group A sensors deliver slightly superior methane quantification precision compared with Group B, yet Group B features far narrower confidence intervals (Group A: 43.5% [95% CI: 34.9, 52.2]; Group B: 45.8% [95% CI: 42.0, 49.7]). These contrasting metrics raise questions about the extent to which cross-team algorithmic choices (alongside wind retrieval biases mentioned above) drive reported performance rather than inherent sensor performance.

Response:

This paper was designed to illuminate the exact complexity the reviewer is highlighting: the interplay between intrinsic satellite sensor characteristics and team-specific algorithmic choices. Our approach is to recognize that both factors are at play and to use their interaction to offer conclusions where they are supported by empirical data. Below are a few areas from the text where we have made changes to address the reviewer's points:

- **When multiple teams using diverse algorithms achieve nearly identical, high performance, it suggests that the sensor's high-quality data is the dominant factor. Regarding Group B performance we write:** *"The seven teams who participated in all four phases of blind testing correctly detected between 28 and 35 of the non-cloudy EnMAP release events and achieved comparable F1 scores of 0.89 to 0.96 (see SI Table S6 for performance metrics). Given the heterogeneity of analysis approaches used, the very similar TP rate suggests some insensitivity to methodological approach and that detection limits are governed by the underlying spectral data."*
- **Conversely, where there is large variability, for example in Group C performance, we use this observation to draw the opposite conclusion: when the sensor provides only marginal signal-to-noise, the analyst's choices become the primary driver of performance. Updated text:** *"The variance suggests that for these instruments operating in their marginal signal-to-noise range, analyst choices like signal thresholding become a dominant factor."*
- **We have revised the abstract to objectively present the Group A and B quantification performance metrics rather than state Group A is outright better. We also note that repurposed instruments perform more variably, but that teams still demonstrate significant skill.** *"Built for purpose" instruments designed to target methane demonstrate excellent detection performance (92-100% true positive rate, 0-11% false positive rate) and quantification accuracy (slope 0.83-1.00, 44% mean absolute percent error, MAPE). Instruments repurposed for methane sensing perform more variably (detection rates 62-88%, MAPE 46%); still, many teams demonstrate significant skill."*
- **We have expanded the discussion around Figure 3, explicitly acknowledging the extreme inter-team variance as an active research area and pointing to large-scale efforts like Project MEDUSA. We also now directly link this problem to a potential solution explored in our own paper: the "wisdom of crowds" ensemble analysis. Updated text:** *"Inter-team variance in quantification is an active research area, with multiple intercomparison efforts underway outside of this study to examine the drivers of high-variance cases (ref 1, ref 2). One strategy for reducing uncertainty may be meta-analyses that combine outputs from multiple teams. Our ensemble analysis shows that aggregating independent estimates can significantly improve quantification precision (see SI S3.9.3)."*

(1) Elderling, 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->.

(2) Akani-Guéry, J., Roger, J. & Buchwitz, M. ESA Project MEDUSA: Algorithm Intercomparison Report (AIR). (2026).

Updated SI S3.9.3: “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).

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.

We then calculated the signed percent error (to assess systematic 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.

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.

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.”

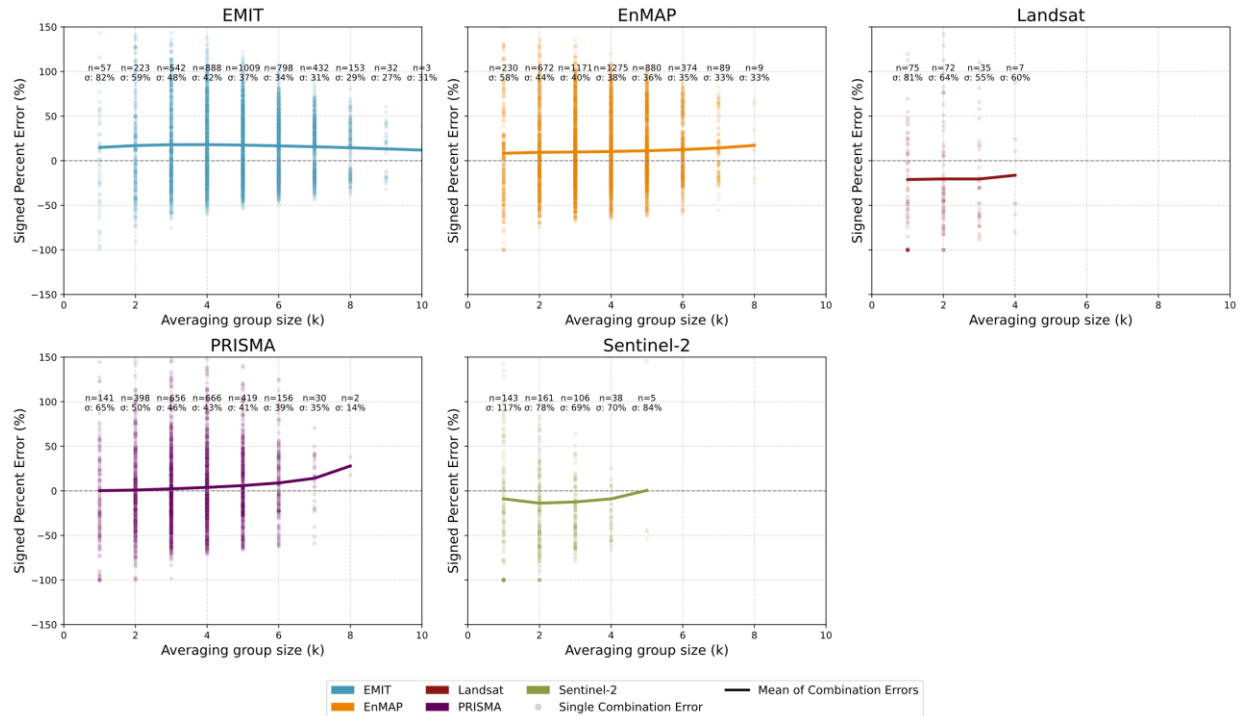


Figure S80. Signed percent error across averaging group sizes k . For each release event, we evaluated the effect of creating an ensemble of k teams, where k ranged from 1 to the maximum number of participating teams (N). For each k , all possible combinations of teams were generated, and the mean of 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.

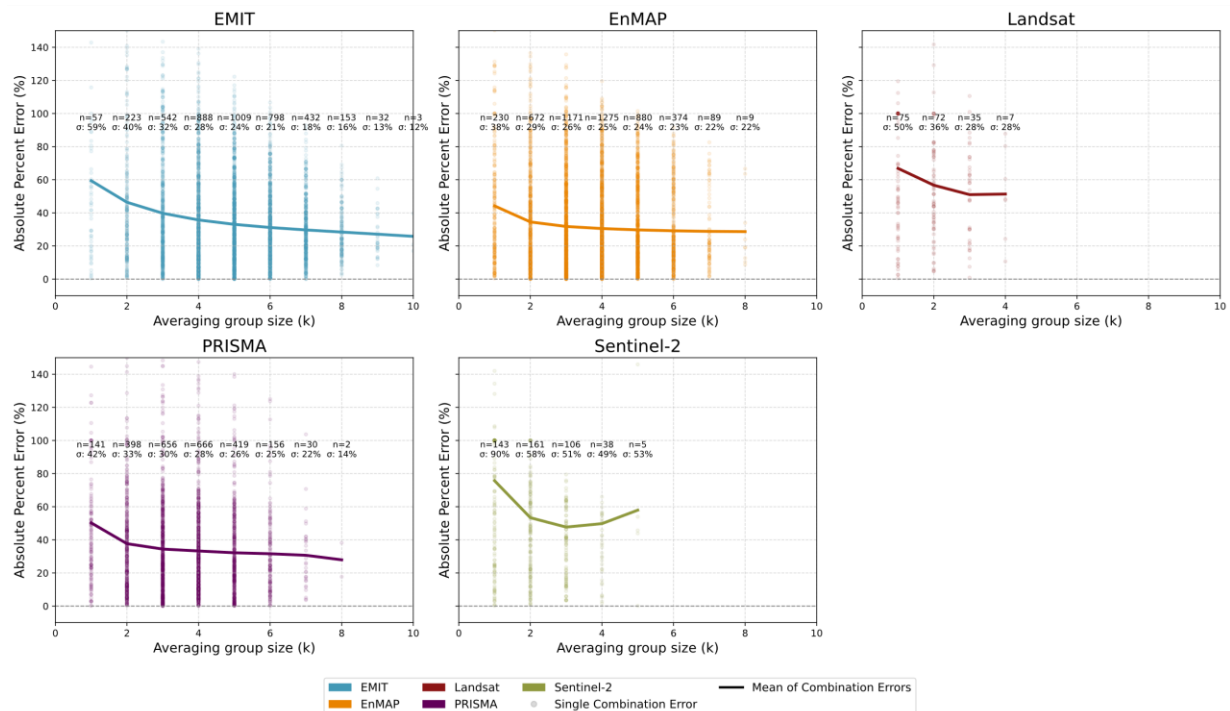


Figure S81. Mean absolute percent error across averaging group sizes k . For each release event, we evaluated the effect of creating an ensemble of k teams, where k ranged from 1 to the maximum number of participating teams (N). For each k , all possible combinations of teams were generated, and the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

Additionally, Lines 276–281 claim that detection results from Group A and Group C sensors show low sensitivity to wind speed, whereas Group B sensors exhibit strong wind dependence. This observation appears contradictory to the narrower confidence intervals measured for Group B instruments.

Response: We have revised the statement to point out weak wind speed effects for Groups A and C and more pronounced but erratic effect for Group B sensors. We’ve also expanded the paragraph to discuss sun angle effects and point out that deeper meteorological exploration is needed outside of this study.

Updated text: “Wind speed and sun angle can also affect plume detection. Group A and C satellites showed weak changes in detection efficacy as a function of wind speed, while Group B showed some sensitivity to winds > 3 m/s, although error bars were fairly wide (see SI Figure S7). Modest decline in TP rate was observed across groups at higher solar zenith angles, consistent with reduced photon availability (see Figure S10). Further examination of meteorological conditions and detection capabilities is warranted for future study.”

Updated SI text: “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.”

(5) Lines 260–270 report weaker correlation between cloud coverage and detection performance for Group A sensors relative to Groups B and C. Figure 4 presents counterintuitive patterns: Group A sensors produce false negatives under near-clear sky conditions, yet exclusively true positives at moderate cloud cover, with zero valid retrievals under 20–40% cloud fraction. For Group B sensors, extensive false positive and false negative detections occur below 40% cloud cover, while zero misdetections are recorded across 11 test cases at 40–60% cloud fraction.

These inconsistent patterns highlight large uncertainties embedded within the evaluation workflow. Beyond ground-based cloud monitoring, all analysis teams should implement scene-level cloud masking algorithms using satellite imagery itself as a standard preprocessing step.

Response: We have created a new, disaggregated version of Figure 4 for the Supplementary Information to provide further context to the main findings between cloud cover and detection performance. This new figure, Figure S11, breaks down detection outcomes by both cloud cover and emission rate bin. 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 True Positives at moderate cloud cover were from higher-rate releases. Similarly, the False Negatives observed for Group B were also predominantly associated with the lowest bin (<250 kg/h) of release size, where the signal is weakest. The "perfect" record seen in the 40-60% cloud bin for Group A, while encouraging, is based on a small sample size.

We have kept the aggregated figure in the main text for simplicity, as disaggregating by both variables leads to sparse sample sizes in many bins.

All participating teams did employ their own scene-level cloud masking algorithms as part of their standard operational workflows. The variation in how they apply these masks is a source of methodological variability. The fact that some teams produce results while others filter the scene is a real component of their performance.

We have updated the text to reference this new figure and clarify these points.

Updated text:

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). 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.

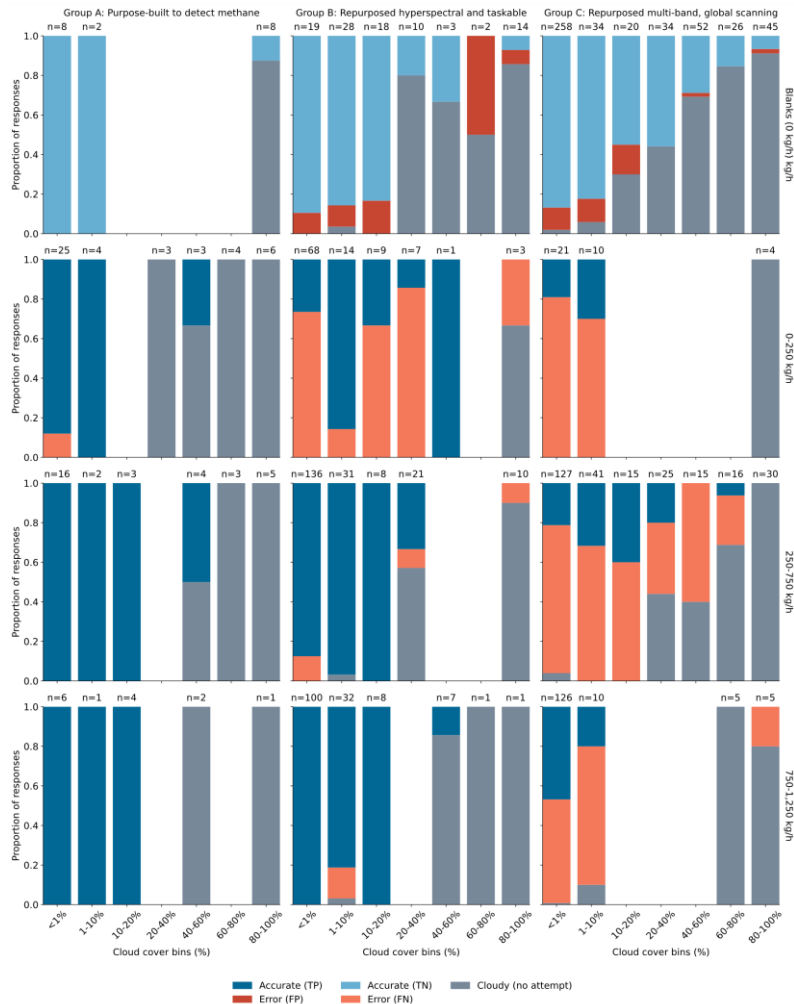


Figure S11. Detection accuracy across satellite typologies, binned by cloud cover percentage and emission rate. Sample per bin listed on chart.

(6) Figure 5 shows that WorldView-3 achieves highly accurate retrievals for low-magnitude methane releases, yet error magnitudes rise substantially for large emission sources; Tanager-1 displays identical behavior. The manuscript should elaborate on the root causes of this trend. One potential contributing factor is the uniform, high-albedo desert surface of the test site. Supplementary Figures S76 and S77 further reveal that the ground release infrastructure occupies a large footprint with strong surface brightness, which likely exerts disproportionate interference on high-spatial-resolution payloads. The highly specialized, homogeneous terrain of this experimental site limits the generalizability of detection performance metrics to real-world methane emission landscapes.

Response: We've updated the text in several places to address the reviewer's multi-layered observations. First, the reviewer astutely points out that WorldView-3 shows high error for large emission sources. Upon re-examining our results, we found that this trend may be characteristic of the analytical method, as this appears to be a broader trend with the

Momentick team analyses seen in Figure S63. It is also plausible the bright ground infrastructure could be a contributing factor. We've updated the Figure S63 discussion accordingly.

Additionally, in the discussion section, we point out that the test-derived quantification fits should not be simply generalized to global emission landscapes. We've updated the text in several places to address the reviewer's points. We also state that future tests should take place in diverse terrains, and that large scale statistical analyses of operational data should complement controlled release testing because they cover a global range of conditions.

Updated SI text: "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.

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 bright footprint of the ground infrastructure interfered with the retrieval."

Updated discussion text: "Ideally, the instrument- and team-specific quantification fits developed here would be useful for characterizing each method's performance and for adjusting estimates of real-world emission sources. However, the applicability of test-based calibration factors to globally measured plumes requires further study. Future validation campaigns should be prioritized to include: 1) locations with varying surface terrains and albedos; 2) locations with different weather conditions and sun availability; and 3) offshore releases to test performance over water.

The experiment design used here, where participants know the approximate source location, best represents a targeted monitoring context where analysts are examining a known area of interest. However, many plumes are discovered through passive, wide-area scanning where no a priori information exists. Therefore, a "fully blind" test that is completely unknown to participants would be the best validation of these survey-level workflows.

While critical, controlled release testing has inherent practical limitations, including environmental impacts and satellite revisit times that constrain sample sizes. Therefore, these

focused validation efforts should be complemented by large-scale statistical analyses of operational data acquired from global surveys. While such “in-the-wild” datasets lack ground truth, they offer vast sample sizes to drive robust performance statistics across a global range of conditions.”

D. Appropriate use of statistics and treatment of uncertainties

No comments.

E. Conclusions

The general conclusions agree with existing consensus on satellite methane sensing, while detailed quantitative results rely heavily on dataset quality and methodological limitations discussed in Section C (Data & methodology).

F. Suggested improvements

All revision recommendations are detailed within Section C (Data & methodology).

G. References

No comments.

H. Clarity and context

(1) Lines 50–60 compare the payloads of Tanager-1 and MethaneSat against TROPOMI. However, multiple hyperspectral sensors with 30 m spatial resolution were launched well before 2024, including Gaofen-5 (2018), PRISMA (2019) and EnMAP (2022). All of these platforms deliver drastically improved performance for methane point-source detection relative to TROPOMI. Therefore, a more relevant benchmark for Tanager-1 would be these peer 30 m hyperspectral instruments, and the manuscript should be revised accordingly.

Response: We’ve updated this sentence to be more inclusive of existing monitoring capabilities besides TROPOMI.

Updated text: “In 2024 a step-change in capability occurred with the launch of two new satellites, Carbon Mapper’s Tanager-1 and Environmental Defense Fund’s MethaneSat. Both systems are public-good missions that were designed specifically to detect methane and release large-scale datasets into the public domain, building on existing monitoring capabilities such as those from high-resolution hyperspectral imagers EnMAP and PRISMA and the coarser-resolution TROPOMI instrument (ref 1).”

(1) Jacob et al 2022

Referee #3 remarks and responses:

Comment 1

I fully appreciate the substantial effort the authors have devoted to addressing the reviewers' comments, with extensive additions and revisions made to the manuscript and supplementary materials that have markedly improved the overall quality of the paper. Nevertheless, as acknowledged by both the reviewers and the authors, this study involves numerous satellites and multiple research teams. While these are inherent features of the work, they also introduce intrinsic limitations: limited and sparsely-distributed samples. Furthermore, algorithms adopted by different teams differ considerably, whereas cross-participation among research teams across different satellites remains extremely low, with only one or two teams performing analyses for some satellites. This inevitably amplifies uncertainties in the results and fails to reach the level of adequate inter-team cross-statistical validation expected for a single-blind test. In addition, methane plumes and emissions are subject to a wide range of influencing factors, including surface conditions, facility properties, wind speed, cloud cover, and solar-viewing geometry. All of the above factors create substantial complexity both for the authors in presenting their findings and for readers in interpreting the conclusions.

Author response: We've added an acknowledgement of the mixed availability of samples and variable cross-team, cross-satellite participation.

Updated methods text: "Teams were permitted to analyze and submit results for any of the 12 satellite observing platforms for which they had access to data. In many cases, multiple teams processed data from the same satellite using independent algorithms and workflows. In other cases, only one or two teams performed analysis for some satellites, limiting total available samples and the opportunity for cross-team validation."

Comment 2

While reviewing this manuscript, I am impressed by the rich and diverse dataset and results presented in this work, yet I am left with substantial conceptual ambiguities following the concerns outlined above. Even for satellites analysed by a relatively large number of groups such as EnMAP, PRISMA and EMIT, the test results show considerable dispersion: some deviate far from the ideal 1:1 reference line, while others align closely with it. This creates significant confusion during my review process. It remains unclear whether better outcomes would be achieved with more extensive inter-team cross-validation or, conversely, with fewer participating teams. As noted by other reviewers and the authors themselves, owing to complexity, sample sparsity and other confounding factors, questions remain regarding the reliability of derived metrics such as minimum detection limits. There are open questions concerning the real-world detection capability of various satellite systems. For instance, WorldView-3 yields a striking minimum detection value of 9.5 kg/h — does this indicate that it delivers the best detection performance? This raises a fundamental question: is a single-blind experimental design inherently ill-suited for multi-target, multi-satellite campaigns that demand long-term execution and heavy logistical support? As the authors mention, it is difficult to achieve sufficient cross-comparison across satellites and research teams and to build a large, dense sample dataset. Without such data, statistical outputs cannot converge toward ground truth, which weakens the clarity and confidence of final conclusions.

Author response: We appreciate the reviewer's reflections about the challenges inherent to this type of research. We have added discussion of these points in the paper and think they are valid questions to be debated in scientific arenas in the coming years. It is heartening to see that controlled releases performed in this study are forming the basis for additional analysis on detection limits (reference 1, GHGSat) and quantification intercomparisons by the European MEDUSA group (reference 2).

(1) <https://egusphere.copernicus.org/preprints/2026/egusphere-2026-3662/>

(2) chrome-

extension://efaidnbmnnnibpcajpcgiclfndmkaj/https://climate.esa.int/media/documents/MEDUSA_D3.2_AIR_MAIN_ANNEX_ABC_v1.0_signed_PUBLIC.pdf?ct=1787085988178

(3) <https://pubs.acs.org/esthag/article/60/21/14961/5107706/Joint-Estimation-of-the-Global-Facility-Scale-Oil>

Samples from main text: “Inter-team quantification variance is an active research area, with multiple intercomparison efforts examining the drivers of high-variance cases (2). One strategy for reducing uncertainty may be meta-analyses combining outputs across teams; our ensemble analysis shows that aggregating independent estimates can significantly improve quantification precision (see SI S3.9.3). “

...

“While critical, controlled release testing has practical limitations, including cost, environmental impacts, and satellite revisit times that constrain sample sizes. Focused validation efforts should therefore be complemented by large-scale statistical analyses of operational data acquired from global surveys. Such “in-the-wild” datasets lack ground truth but offer vast sample sizes to drive robust performance statistics across a global range of conditions (ref 3)”

Comment 3

Regarding the Main section at the beginning of the manuscript: based on the overall content and conclusions, this study is not primarily focused on Tanager-1 and MethaneSat; indeed, MethaneSat ultimately provided no usable data for this work. In essence, this paper reports a comprehensive single-blind methane-detection benchmark covering currently operational ultra-high-spectral-resolution satellites (spectral resolution < 1 nm), high-spectral-resolution satellites (spectral resolution > 1 nm), high-spatial-resolution satellites, and wide-swath multispectral satellites. This feature should be explicitly stated in the text.

Author response: We’ve added a new paragraph to more holistically discuss the broad classes of satellites covered in this work.

“Several broad classes of instruments have been brought to bear on methane detection¹. Purpose-built point-source sensors began with GHGSat-D in 2016, which has since expanded into a GHGSat constellation of more than ten satellites offering greenhouse gas monitoring as a service¹⁵. In 2024, two new public-good missions launched specifically for methane detection—Carbon Mapper’s Tanager-1 and Environmental Defense Fund’s MethaneSat. Tanager-1 is a high-resolution

hyperspectral point-source sensor, while MethaneSat, operational until June 2025, had finer spectral resolution and coarser spatial resolution, targeting wide-area total emissions estimates^{16,17}. Both missions build on the capabilities of a large group of multi-purpose, high-resolution (<60m) hyperspectral instruments including EnMAP, EMIT, Gaofen 5, PRISMA, and the coarser-resolution TROPOMI¹. Higher-spatial-resolution (<10m) commercial systems include GHOSat and Worldview-3^{12,18}. Another category of multispectral land-scanning satellites, such as Sentinel-2 and Landsat, have wide swaths that provide regular revisits and global coverage.”

Comment 4

Moreover, I suggest a more objective description of the international development timeline for hyperspectral methane-monitoring satellites, from Gaofen-5 (2018), PRISMA (2019), EnMAP (2022), EMIT (2022), through to Tanager-1 (2024). These hyperspectral instruments belong to fundamentally the same category of payloads; therefore, the term "step-change" may be inaccurate.

Author response: We’ve removed the characterization of a ‘step change’ and expanded the description of other hyperspectral and methane monitoring payloads.

“Several broad classes of instruments have been brought to bear on methane detection¹. Purpose-built point-source sensors began with GHGSat-D in 2016, which has since expanded into a GHGSat constellation of more than ten satellites offering greenhouse gas monitoring as a service¹⁵. In 2024, two new public-good missions launched specifically for methane detection—Carbon Mapper’s Tanager-1 and Environmental Defense Fund’s MethaneSat. Tanager-1 is a high-resolution hyperspectral point-source sensor, while MethaneSat, operational until June 2025, had finer spectral resolution and coarser spatial resolution, targeting wide-area total emissions estimates^{16,17}. Both missions build on the capabilities of a large group of multi-purpose, high-resolution (<60m) hyperspectral instruments including EnMAP, EMIT, Gaofen 5, PRISMA, and the coarser-resolution TROPOMI¹. Higher-spatial-resolution (<10m) commercial systems include GHOSat and Worldview-3^{12,18}. Another category of multispectral land-scanning satellites, such as Sentinel-2 and Landsat, have wide swaths that provide regular revisits and global coverage.”

Comment 5

Following the launch of Gaofen-5, Ziyuan-1 and PRISMA, dozens of international research groups conducted methane-detection campaigns over the Permian Basin in the United States. These studies first pushed the detection limit for point-source

methane emissions down to the hundreds-of-kilograms-per-hour level. Key findings were originally reported in Science Advances (Itziar Irakulis-Loitxate et al., 2021). Strictly speaking, these payloads were not "repurposed towards methane observation".

Author response: We've removed the word 'repurposed' in three places in the main text.

"Group B instruments are ~~repurposed~~ hyperspectral and/or high spatial resolution instruments that were originally launched for broader observing purposes but are used to detect methane."

"New satellites with diverse capabilities have been launched, while other existing multi-purpose systems have been ~~repurposed~~ applied towards methane observation"

"We have shown that purpose-built sensors can robustly detect methane emissions near 50-200 kg/h under ideal conditions, and many teams can also wield ~~repurposed~~ multi-purpose instruments to detect and quantify methane with significant skill."

Comment 6

Concerning the scarcity of Gaofen-5 and Ziyuan-1 data obtained in this experiment: the authors attribute this mainly to "Logistical & geopolitical barriers". I have consulted the operational authority for Gaofen-5, which demonstrates full willingness to participate in and support this valuable scientific experiment. Substantial funding and manpower from China have been invested in Gaofen-5 and Ziyuan-1 precisely to address these major scientific challenges facing human society. Admittedly, unfamiliarity and differing team-selection practices across national-regional collaborations have created practical communication and cooperation difficulties.

Author response: We share the reviewer's disappointment that we couldn't obtain more datapoints for GF5 and ZY1. Between April and November 2025, we emailed our contact for these instruments on a near monthly cadence requesting advanced notice of observation opportunities to coordinate releases (samples below). Their replies indicated support and willingness to participate. However, unfortunately, this rarely translated into advanced notice about overpass opportunities. We've added a sentence to the Methods noting this.

Methods text: "The experiment team tracked satellite observation opportunities using a combination of email communications with participating teams, online user portals, and predictable orbital schedules. The scheduling of observations was determined by each satellite operator according to their own mission priorities and commercial considerations and was out of the control of the experiment team. Unfortunately, only a few advanced notice observation opportunities were offered with Gaofen 5 and Ziyuan-1. For instruments requiring tasking, we generally requested as many observations as

were available. Exceptions include requests for EnMAP and PRISMA, which were occasionally modulated to manage total volume, and for GHGSat, which provided a

Comment 7

For Figure S1 in the new supplementary materials: I recommend removing the "min. test rate" column. The minimum-detection-limit value of 9.5 kg/h derived from WorldView-3 reaches performance levels comparable to airborne instruments. Yet the authors acknowledge in their Response that such figures may be biased by high surface reflectance and other artifacts. Retaining these values in the figure could easily mislead readers.

Author response: We understand how the minimum test rate column could cause confusion. We've removed it from Figure 1 and Figure S1.