Single-Blind Testing of Continuous Monitoring System With Realistic Emission Timeseries

David Ball[1]

[1]Project Canary

October 2024

Abstract

In this report, we assess Project Canary's performance during a single-blind trial of a sophisticated controlled release experiment ("ADED 2.0") performed by METEC intended to more accurately mimic emissions at oil and gas facilities via the inclusion of a noisy and time varying background layered with large asynchronous releases. Even under these more complex testing conditions, we see good agreement between Project Canary's quantification output and the ground truth rates: at the end of the 28 days of blind testing, Project Canary reported that 674 kilograms of methane had been released, when the facility had actually released 701 kilograms, representing a cumulative underestimation of 3.8% and a mean rate error of -0.04 kg/hr. Despite the increased difficulty associated with the significantly more complex emissions patterns of this overhauled testing protocol, the key error metrics are consistent with Project Canary's latest publicly-available results from the ADED 2024 campaign. Finally, we investigate the impact of sensor density by applying the same quantification algorithm to data from only 3 of the 10 sensors deployed at METEC and find that the facility-level quantification accuracy is not significantly impacted by a reduced sensor count.

1 Introduction

There is a clear need in the methane measurement industry for robust third-party blind controlled release testing. With so many competing technologies, each with different performance trade offs and costs, it is essential to have a set of standardized tests that the technologies can be benchmarked against in order to assess their relative performance across a variety of metrics spanning detection statistics, localization, and quantification accuracy. Not only do these blind tests provide standardized evaluative statistics, but perhaps more importantly, they provide large volumes of high-quality "ground truth" data that technology developers can use to drive innovation and improve their systems.

Historically, the "Advancing Development of Emissions Detection" (ADED) protocol run by the Methane Emissions Technology Evaluation Center (METEC) at Colorado State University has been the standardized annual test that the majority of Continuous Monitoring System (CMS) technologies participate in. The ADED protocol was run for three consecutive years (2022, 2023, 2024) and used to assess the performance of continuous monitoring systems through time ([1, 2]), which generally have improved year-over-year in terms of several accuracy metrics.

While the results from the most recent iteration of ADED (i.e., ADED 2024) are not yet published in a comparative paper, every solution's report detailing key metrics are publicly available on METEC's website (`https://metec.colostate.edu/aded-testing-results/`). Over the course of these testing campaigns, Project Canary demonstrated significant improvements across metrics spanning detection statistics, localization, and quantification accuracy. While a more detailed discussion of ADED 2024 performance is outside the scope of this writeup, we provide here a brief summary of Project Canary's ADED performance and associated key metrics from the 2023 and 2024 tests. We note first that these improvements were entirely algorithmic: from improving data preprocessing and cleaning, to more accurate dispersion modeling, and finally implementing more sophisticated inverse solvers. In fact, the hardware that collected the data was identical between 2023 and 2024. The first and perhaps most notable measured improvement in 2024 is that Project Canary significantly improved its detection

statistics. More specifically, the false negative fraction decreased from 27% to 8% with only a marginal increase in the false positive fraction (which went from 10.6% to 14%, driven by slightly overcounting leaks during a positively-identified emission event, **not** from reporting emissions when none were present on the facility). This resulted in a dramatically lower 90% Probability-of-Detection (POD), which decreased from 6.2 kg/hr to 0.5 kg/hr between 2023 and 2024 (an improvement by a factor of 12, and the lowest-ever recorded 90% POD during an ADED campaign by a factor of 7). Project Canary's performance in terms of source localization also improved: in 2023 we accurately localized to the equipment unit 40.8% of the time and to the correct group 53.9% of the time (for a total of 94.7% of releases being localized to at **least** the correct group). In 2024, 50.4% of the releases were attributed to the correct unit (an improvement of about 10%), and 98.8% of releases were localized to at **least** the correct group (an improvement of 4.1%). Finally, the quantification estimates also improved: in 2023 the mean quantification error and factor of rates that were correct to within a factor of 2 were: -0.345 kg/h and 60%, respectively, while in 2024 these metrics improved to -0.064 kg/hr and 75%.

To expand beyond the evaluative statistics provided directly by METEC in their reports, we also compute the cumulative emissions error (in total kg) over the 90-day testing periods (an error metric that we believe is of high import given recent OOOOb regulations and requirements for CMS). In 2023, METEC emitted 1,969 kilograms of methane during the roughly 90 day testing period and Project Canary estimated a total of 1437 kilograms released from the facility: an underestimation by nearly 30%. In 2024, METEC emitted 2,272 kilograms and Project Canary estimated a total release mass of 2,208 kg, an error of less than 3%, representing an improvement in the cumulative emissions accuracy by a factor of 10. These results are summarized below in Table 1

Table 1: Evaluative Metrics from Project Canary's 2023 and 2024 ADED Results

Metric	PC 2023	PC 2024
False Negative Fraction	27%	8%
False Positive Fraction	10.6%	14%
90% Powerlaw POD (kg/hr)	6.2	0.5
Localization Precision (Equipment Unit)	40.8%	50.4%
Localization Precision (Equipment Unit + Group)	94.7 %	98.8%
Mean Quantification Error (kg/hr)	-0.345	-0.064
Percentage of Rates within Factor of 2	60%	75%
Cumulative Emissions Percent Error	-27%	-2.8%

When considering these improvements and how to interpret them in the context of a relatively simple annual testing where the protocol remains unchanged, the astute observer may be somewhat skeptical of how these results will generalize to more realistic scenarios. More specifically, the original ADED protocol is quite simple in terms of its emissions patterns and what it asks the solutions to provide: the three-month test is broken up into individual "experiments", each of which has between 1 and 5 individual releases. The individual release rates are held constant during each experiment and are turned on and off simultaneously at the experiment's start/end times. As such, the facility is either in a sterile "off" state with no emissions, or an "on" state with constant rates. These relatively simple and known patterns effectively represent "prior" information that the solutions can, in principle, leverage in their algorithms. Furthermore, the experimental design forces solutions to employ an "event-based" quantification reporting paradigm, which is at odds with how an optimal system would work in the field. In other words, the problem is much more algorithmically challenging if source rates cannot be assumed to be constant and if individual sources can turn off and on asynchronously, which is expected at operational sites. In light of these relatively simple and known patterns combined with the abundance of historical data to lean on, the skeptical observer may suspect that any demonstrated improvements in performance could be due to the solutions over-optimizing their systems to the relatively simple patterns of the testing protocol. If this were the case, then the performance demonstrated during ADED testing would not generalize to more complex emission patterns that are expected in the field.

To address the valid and pressing concerns mentioned above, the METEC team has been working to implement an advanced testing protocol designed to more accurately mimic the complex emissions present at operational facilities. In Section 2, we describe some of the details of this new testing

protocol and contrast its features with the original ADED protocol. Section 3 presents Project Canary's quantification estimates during a blind 4-week test of this new controlled release program (henceforth referred to as "ADED 2.0").

2 ADED 2.0

While we at Project Canary are not privy to every detail of how the ADED 2.0 emissions timeseries are generated, we can speak to some of the general features evident in the ground-truth data that was provided to us after a blind four-week trial period that we participated in (more on this in Section 3). The description here represents our best understanding of these tests and does not come directly from METEC, for any official explanation or inquiries, please contact them.

The goal of this overhauled testing protocol is to more accurately mimic the emission timeseries that are expected at an operational facility. This includes a noisy time-varying background with significant high-frequency power from a variety of different locations, representing operational emissions (e.g., pneumatic devices, compressor slip, etc.). After a 1-week period of only "baseline" emissions (intended to be representative of a baselining period defined by OOOOb regulations), METEC begins to add larger releases of varying rates and durations. These larger releases may overlap in time entirely, partially, or not at all (in contrast to the previous ADED protocol where simultaneous releases always start and stop at the same times). Figure 1 shows METEC's releases during the four-week trial period. In this figure, each color corresponds to a different equipment group, and they are stacked on top of one another such that the height of the stacked colors corresponds to the source-integrated (i.e., facility-level) emission rate. All of the previously-described complex features of this updated release protocol are evident in this figure: there are high-frequency noisy emissions from every group at the facility along with low-level baseline emissions, layered with asynchronously-occurring larger emissions.

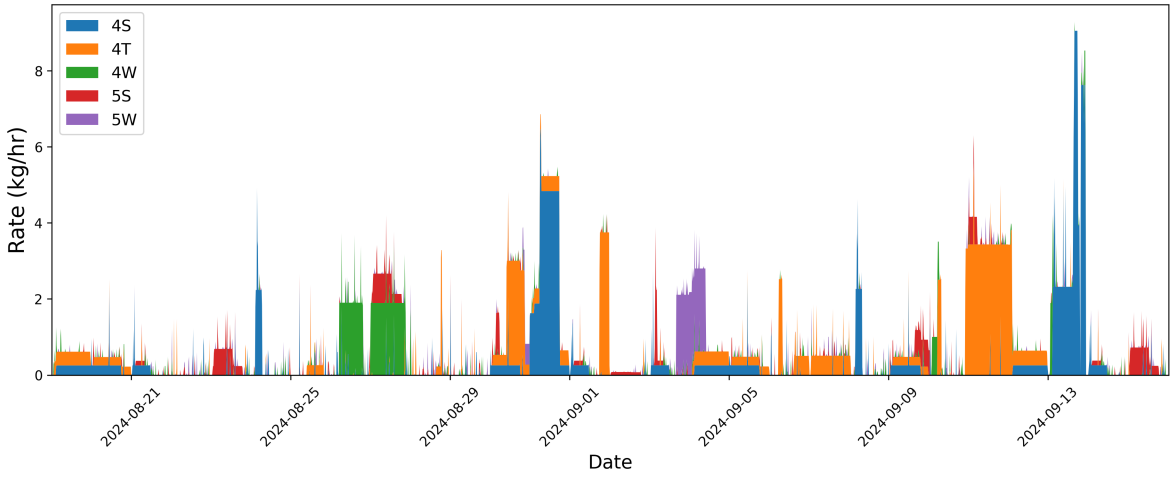

Figure 1: Ground-truth emission rates during the entire 4-weeks of the ADED 2.0 trial held in late August and early September of 2024.

To illustrate the contrast between the original ADED protocol and ADED 2.0, we show in Figure 2 the emission timeseries from 250 hours (a zoomed-in snapshot) during the ADED 2024 campaign (top) and ADED 2.0 trial (bottom). Comparing these timeseries, it is immediately evident that there is significantly more complexity in the ADED 2.0 releases as compared to ADED.

In the original ADED protocol, the event-based nature of the facility's emission leads to the algorithms almost writing themselves: first detect the start/end times of a given emission event based on some fairly simple features that are evident in the methane measurement signals, and next infer the best-fit source rate of a given event from every equipment unit on the facility, assuming a constant rate over the event's duration (of course there is a lot of detail in the event detection, localization, and quantification steps, I don't mean to imply that those are trivial to implement and get good results, but the general framework is relatively simple). When thinking about how to generalize these algorithms to more complex emission scenarios, it becomes immediately obvious that an event-based framework

will not suffice: in ADED 2.0, there is no such thing as a cleanly-defined "emission event", which are a key component of the original ADED protocol. The implication of this is that the algorithms to infer source rates must take into account the fact that emission rates may change dramatically on short timescales and also start and stop asynchronously: a continuous estimator is needed.

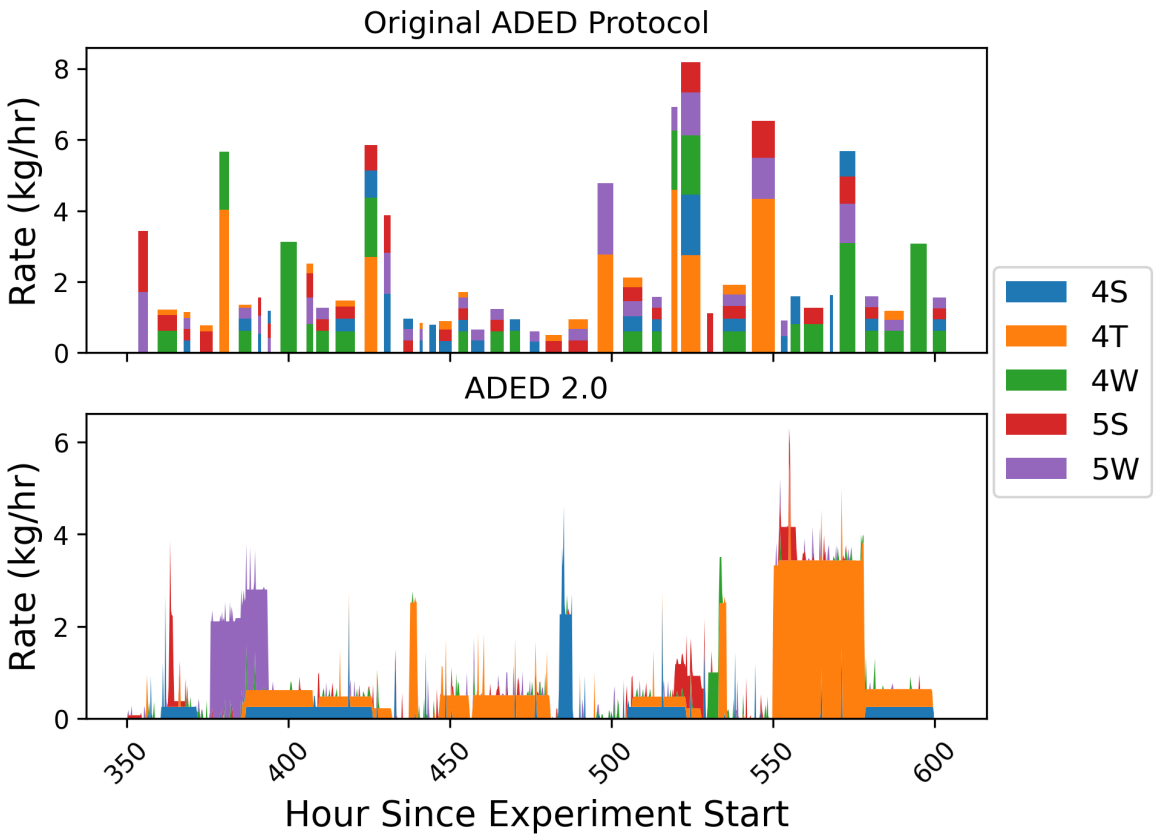

Figure 2: Ground-truth emission rates during arbitrarily-selected 250 hours from the original ADED protocol (top) and ADED 2.0 (bottom). Each color corresponds to a specific equipment group.

3 Project Canary's Results During a 4-Week Trial of ADED 2.0

Project Canary was invited to participate in a completely blind 4-week-long trial of this new protocol (along with a few other major continuous monitoring technologies). Not only was this testing blind, but this was the first test of its kind. In other words, while for ADED, there may be concerns about technologies over-optimizing algorithms to the testing protocol, there was no possible way we could leverage any past controlled release testing to optimize for this blind trial, because no testing data such as this exists. As such, this was the best possible test at METEC of an out-of-the-box deployment of our system. Project Canary has invested heavily in the last couple years developing more sophisticated continuous estimators (rather than event-based quantification schemes), exactly because we recognized the dramatic differences between the simplified testing protocols and the expected emissions in the field. For this reason, we were confident that our solution, without any specific tuning, would perform well even under the complex emission scenarios present in this more advanced testing protocol.

Figure 3 shows a side-by-side comparison of the releases during the 4-week-long ADED 2.0 trial (top panel) and our blindly estimated rates, which were reported in real-time every 15 minutes in our web-based dashboard and downloaded directly by the testing center for analysis. Visually inspecting these rates, we see a fairly close correspondence between the facility-aggregated quantification estimates and the ground-truth release rates and generally good correspondence between the colors (i.e., localization)

between the two plots (in other words, when there is a large emission from a given equipment group, it is generally reflected in our quantification estimate). We note that the correspondence is not perfect, nor do we expect it to be: in general we have seen from years of ADED testing, even under simple emission scenarios, that there is significant scatter in the error distributions of quantification estimates from continuous monitors, however the systems can achieve very low bias (see the "mean quantification error" statistic in Table 1). The implication of this is as follows: short-duration rate estimates from continuous monitoring systems are prone to significant error, but over longer aggregation times can yield an accurate accounting of emissions due to the low bias of the system. The error distribution of these 15-minute reported quantification estimates is shown in Figure 4. While the instantaneous 15-minute errors have some scatter about 0 (for reference, the mean absolute error is 0.66 kg/hr), the distribution is centered close to 0, with a mean error (i.e., bias) of -0.04 kg/hr. In short, any single instantaneous rate estimate should not be taken too seriously, but time-aggregated (e.g., daily, weekly, monthly) estimates show a high degree of accuracy due to the low bias of the system.

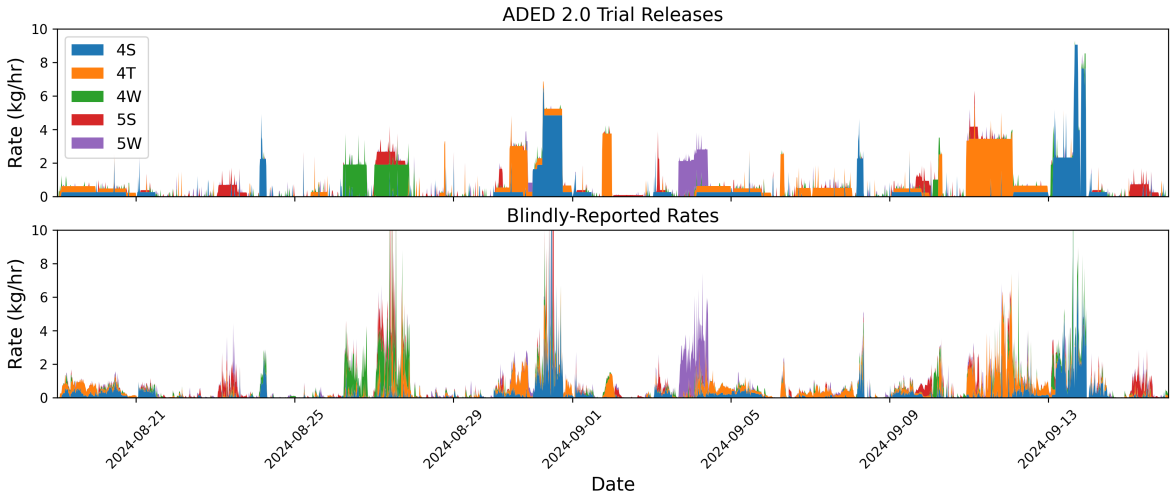

Figure 3: Ground-truth release rates (top) and blindly-estimated rates (bottom) reported every 15 minutes.

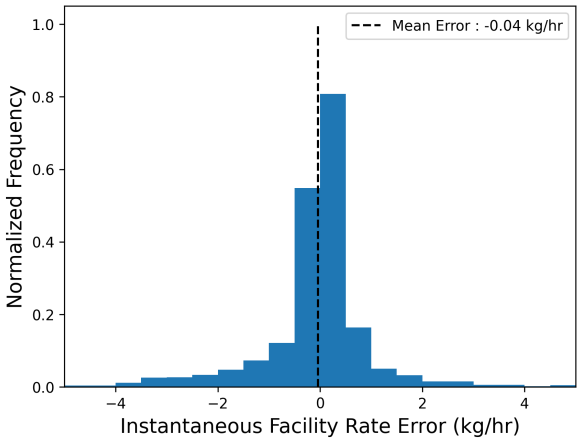

Figure 4: Error distribution of facility-level quantification estimates during the blind ADED 2.0 trial. While not every 15-minute quantification estimate is highly accurate (although the majority are within 1 kg/hr), the bias is very close to 0, indicating that the system's output, when integrated over long time periods, is highly accurate.

To illustrate the implications of the low-bias nature of the system, we compute 12-hour averages of the ground-truth release rates as well as our blindly-reported estimates. The specific choice of 12 hours as the aggregation time is loosely motivated by the EPA's requirement under the OOOOb Continuous Monitoring Alternative Test Method that a continuous monitoring system reports a quantified emission rate from the facility at least once every 12 hours. As such, we thought it would be prudent to visualize the 12-hour means of both the blindly-estimated and actual rates, shown in Figure 5. When considering these longer averaging times, the correspondence between the actual rates and estimated rates is even more striking. While there is still some error (some of the peaks are slightly overestimated, and some are underestimated) and source confusion (the blindly-reported rates show the slight tendency to overproduce nonzero sources), the visual correspondence between the two panels is encouraging.

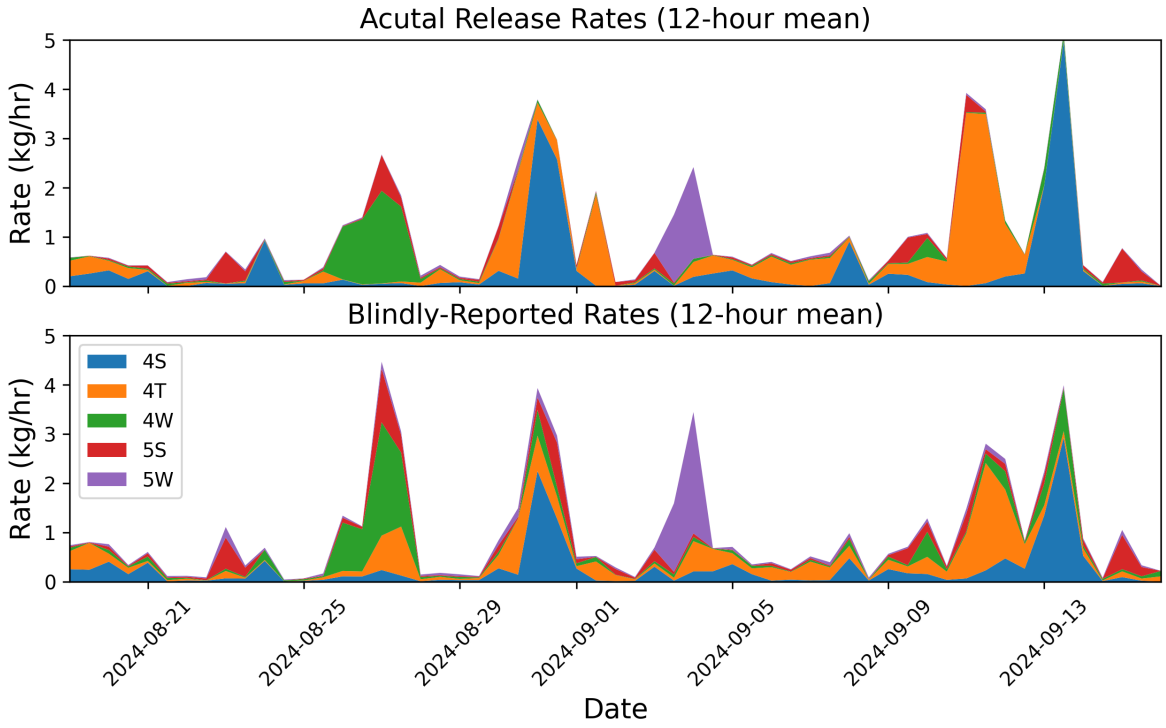

Figure 5: 12-hour aggregated mean release rates from the facility (top) and blindly-estimated 12-hour aggregated mean rates by Project Canary's system (bottom)

In order to more quantitatively demonstrate how the low bias of the instantaneous rate estimate results in a highly-accurate cumulative accounting of methane emissions over long timescales, we show in Figure 6 the cumulative emissions curves of the blindly-reported rates compared to the actual cumulative emissions through time from the facility. We see the same general trend that is evident in the error distribution histogram: while the system can often under or over predict a given short release event, it does so in equal measure, resulting in an accurate estimate of total facility emissions over long timescales: at the end of the 28 day blind testing period, Project Canary's system had reported an estimated total mass emission of 674 kilograms, while the facility had actually released 701, an underestimation by only 3.86 percent.

3.1 Dependence on Sensor Density

One common and valid critique of extrapolating controlled release testing results (specifically at METEC) to the field has to do with the number of sensors technology providers tend to deploy during these tests. Historically, point-sensing solutions have deployed between 8 and 12 sensors at METEC during the ADED campaigns (the exact number deployed by each solution is published in their respective reports), but only suggest deploying between 3 and 5 sensors in practice. As such, the results from METEC represent a "best-case scenario" in terms of a system's performance due to the higher density of information gathered by the denser sensor network. While we cannot speak for the

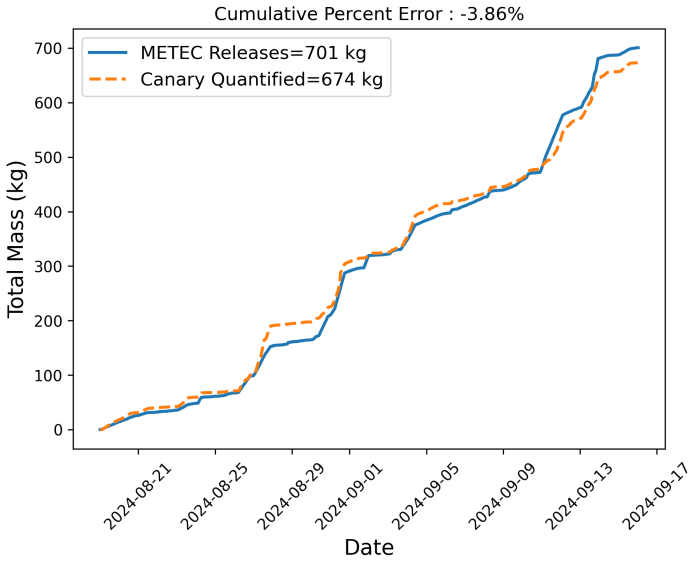

Figure 6: Cumulative emissions curves of the blindly-reported rate estimates (dashed orange) and actual rates (solid blue). At the end of the 4 week testing period, Project Canary had blindly reported 674 kilograms of methane being released from the facility, an underestimation of the actual value by 3.86 percent.

other companies tested at METEC, our reasons for deploying an overabundance of sensors at controlled release testing facilities is twofold: first, and most importantly, this controlled release testing data is incredibly valuable to us in terms of building and refining models and algorithms, and the more data we have during time periods with known ground-truth source rates, the more we are able to improve our understanding of the dispersion of methane under various conditions. In other words, more data is better, and the METEC releases are a source of extremely high-fidelity and valuable controlled release data. The second reason is simply that virtually all of the solution providers deploy an overabundance of sensors, and the sensor count is often swept under the rug during discussions and analyses of these results. As such, deploying fewer sensors may put a participant at a competitive disadvantage in terms of the optics of how these results are interpreted.

We are keenly aware of this significant discrepancy between controlled release testing setups and deployment in the field, and specifically design our algorithms to be as robust as possible against varying sensor density. While there will invariably be certain performance hits when deploying a sparser sensor network, we take care to make sure that the most important metric to us (specifically the cumulative quantification error, driven by a low-bias system) is roughly invariant to changes in sensor count. In order to demonstrate this, we apply the exact same quantification algorithm to the measurement data taken during ADED 2.0 from only three sensors. For this reanalysis, we pick the three sensors that most closely match what we would deploy at an operational facility when considering the potential source locations and prevailing wind directions over the year. These results are shown in Figures 7 and 8 which show the 12-hour average rates broken down by equipment group and the cumulative emission curves, for the actual releases, the blindly-reported results (using 10 sensors), and a recalculation of rates using exactly the same quantification algorithm but only a subset of the input data (from 3 sensors). Considering Figure 7, we see only minimal differences in the 12-hour means between the 10-sensor quantification output and 3-sensor output. More specifically, the 3-sensor output is prone to slightly more source confusion (i.e., the localization accuracy is affected, an expected impact of reducing the sensor count), but the overall key features of the emissions curves are preserved (dominant source and facility-integrated rates). Figure 8 shows the cumulative emission curves for the actual emissions alongside the previously-described 10-sensor output (i.e., the blindly-reported rates) and also the quantification output of the three sensor network. While differences are noticeable in the cumulative emissions between the 10 and 3 sensor curves, these deviations are minimal: the cumulative emissions estimate at the end of the 28-day testing period only differed by 1 kilogram between the 10 and 3 count sensor networks. Ultimately we see that while some secondary metrics will be negatively

affected by a reduced sensor count, the primary metric of import (the site-level cumulative emissions accuracy) is preserved even when reducing the sensor count from 10 to 3 sensors.

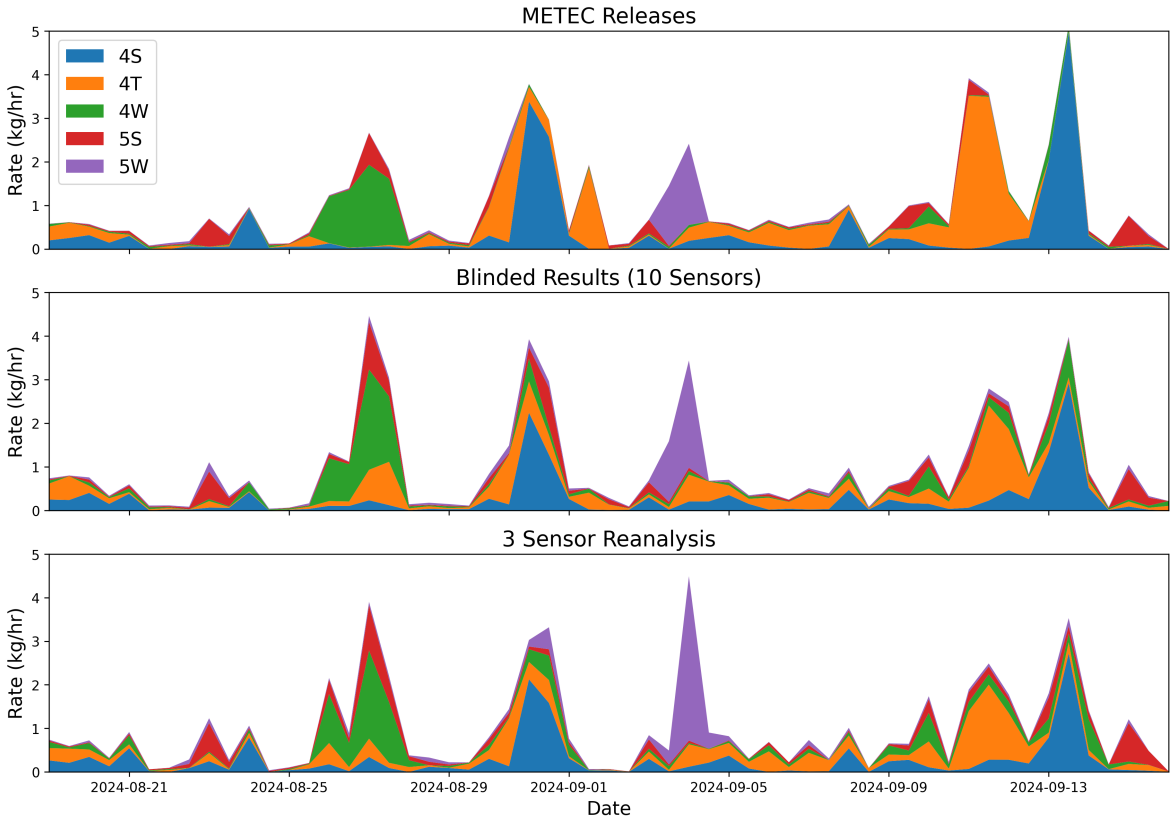

Figure 7: 12-hour means of actual release rates (top), blindly estimated rates with 10 sensors (middle) and 3-sensor quantification output (bottom).

4 Conclusions

In this report, we began by looking at some of the significant progress that has been realized via algorithmic and modeling improvements year-over-year by Project Canary during the original ADED testing campaigns (2022-2024) and the associated key metrics. While these improvements are encouraging, we mention several reasons to be skeptical of how generalizable these results are to the field, in light of the repeated and oversimplified testing protocol that does not fully capture the complexities of emissions patterns expected at operational oil and gas facilities. We then describe a new testing protocol being developed by METEC meant to directly address exactly these limitations by incorporating more realistic emission patterns, including noisy and time-variable baselines as well as asynchronously starting-and-stopping larger releases. We present results from a blind controlled-release study meant to serve as a trial run of this new protocol that lasted for 4 weeks. Generally, there is very good agreement between Project Canary's blindly reported results and the ground truth releases: the dominant emission source locations and facility-integrated rates correspond well, and the cumulative emissions error at the end of the 28 days was only -3.86%. Despite the significantly more complex emissions patterns employed in this blind controlled release study, the key error metrics are consistent with Project Canary's ADED 2024 results. Finally, we recompute the quantification estimates with a reduced sensor count network and demonstrate that the key metric of import (the cumulative site-level emissions error, which is dictated by the system's bias) is invariant to changing the number of sensors from 10 to 3, showing that even under complex emissions patterns and reduced sensor count, both of which are expected in the field and are sources of concern and skepticism about extrapolating results from controlled release studies to the field, that the system's bias remained nearly 0, resulting in a

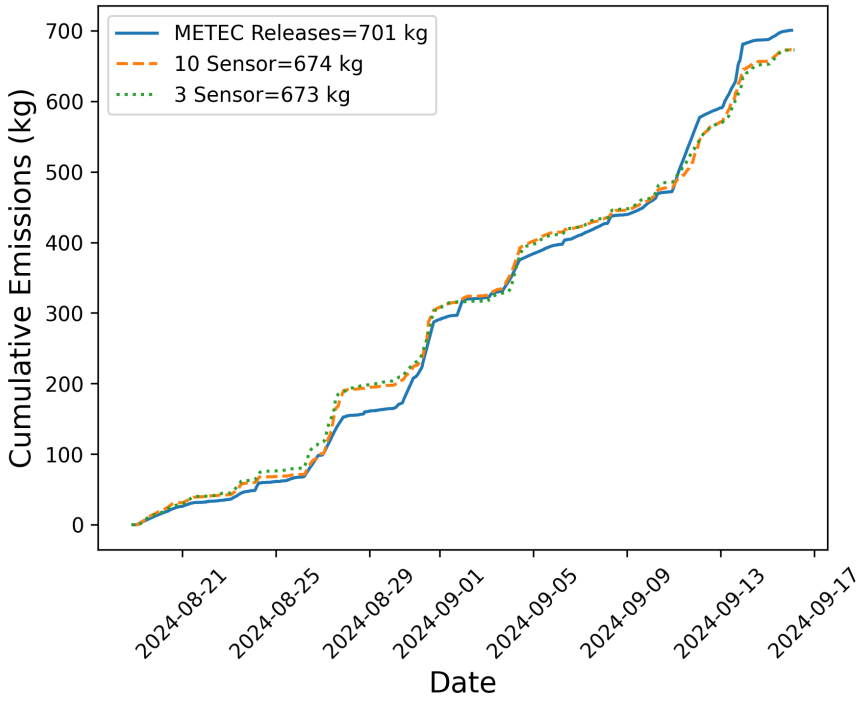

Figure 8: Cumulative emissions curves of the blindly-reported rate estimates (dashed orange), actual rates (solid blue) and recomputed quantification estimates using only three sensors (dotted green). Only minimal differences are present in the cumulative emissions estimates of the 10 and 3 sensor networks.

highly-accurate accounting of the methane emissions over the 28 day blind test, even with only 3 sensors' input data.

We believe that these results represent a significant step forward in demonstrating what continuous monitoring systems are able to achieve in terms of quantification accuracy of total site-level emissions, even under complex emission patterns and realistic sensor deployment in terms of sensor count and positioning. In light of the EPA's recent requirements for continuous monitoring as an Alternative Test Method, the application of CMS as addressing an accounting problem (i.e., "tell me how much mass was emitted over this 7 or 90 day period") as opposed to a leak detection problem (i.e., "did you detect and accurately quantify a short-duration event?") highlights the importance of a particular key metric: the bias of quantification system, which can be directly measured via the cumulative site-level emissions estimate during controlled release tests. We demonstrate even under complex emissions and realistic sensor density that our system consistently achieves a near-0 bias in the total site-level emissions estimates, resulting in highly accurate accounting of total mass emitted over long timescales.

References

[1] Clay Bell et al. "Performance of Continuous Emission Monitoring Solutions under a Single-Blind Controlled Testing Protocol". In: *Environmental Science & Technology* 57.14 (2023). PMID: 36977200, pp. 5794–5805. DOI: 10.1021/acs.est.2c09235. eprint: https://doi.org/10.1021/acs.est.2c09235. URL: https://doi.org/10.1021/acs.est.2c09235.

[2] Chiemezie Ilonze et al. "Assessing the Progress of the Performance of Continuous Monitoring Solutions under a Single-Blind Controlled Testing Protocol". In: *Environmental Science & Technology* 58.25 (2024). PMID: 38865299, pp. 10941–10955. DOI: 10.1021/acs.est.3c08511. eprint: https://doi.org/10.1021/acs.est.3c08511. URL: https://doi.org/10.1021/acs.est.3c08511.