Earthview Corporation
908 Dragon St
Dallas TX, 75207

Key Takeaways from 2024 METEC ADED Single-Blind Testing

Earthview
2 January 2025

The results discussed below represent the performance of the BluBird system under the test scenarios and environmental conditions encompassed by the 2024 METEC ADED test period. For additional background and results, see white paper blubird_ADED_2024_performance_overview_pub_2024_12.pdf.

Rate of Detection: The Earthview BluBird 2.0 system detected[1] 91% of the 347 individual gas release events ("experiments") during the winter/spring 2024 ADED experiment. This increases to 92%[2] and 94%[3] when alternative criteria are applied that are more flexible regarding reported start times for events (see footnotes).

For ADED experiments consisting of a single leak source, BluBird detected 87% of these events. This increases to 92%[2] and 97%[3] using the alternative start-time and overlap criteria.

Ability to Detect Low Emission Rates: Ninety five percent of single-release experiments with release rates of 1000 g/h or less were detected, based the 30-minute criterion[2]. Eighty-five percent were detected using the 20-minute criterion[3].

These increases in detection count are associated with relatively slight changes in allowable start times (e.g., 30 minutes early vs. 20 minutes early). This suggests that METEC's ADED analyses should consider the degree to which reported results are affected by such fairly minor modifications in their classification technology.

Probability of Detection (POD): POD as a function of leak release rate (flow rate) was calculated using logistic regression applied to binary values of "detected" and "not detected" classifications. Separate calculations were done using the set of single-release experiments and for all experiments. For the latter, which included some experiments with up to 5 simultaneous releases, the maximum flow rate among the releases was used (the single-release events were also included in this set).

While the normal way of thinking about POD assumes that leak rates have the biggest effect on detection rates, this will depend on a number of factors. To investigate this, POD was also calculated as a function of leak duration and as a function of total gas released per experiment.

For the single-release experiments, the estimated 90% POD ranges from 2400 g/h with detections classified using the 20-minute start time threshold to 990 g/h using the 30-minute start time threshold (Figure 2). Allowing the extra 10 minutes added an additional 6 detections due to the list of true positives. (Detection start times were averaged to the nearest minute.).

Earthview Corporation
908 Dragon St
Dallas TX, 75207

When the more lenient overlap criterion is applied (Figure 2), 11 additional true positives were included, resulting in a decrease in 90% POD decreases to 320 g/h.

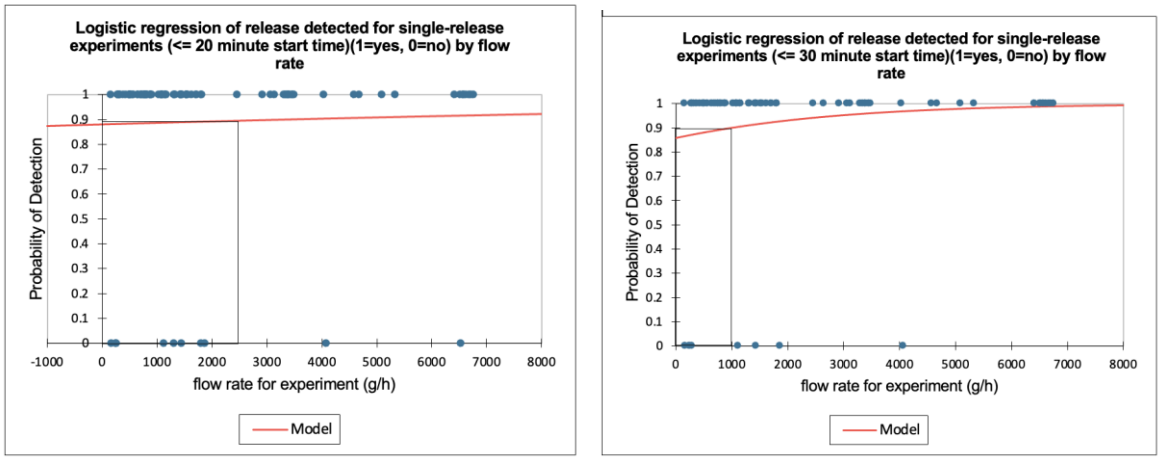


Figure 1. Probability of detection as a function of leak flow rate for experiments consisting of a single release. In the left-hand plot, detection reports meeting the 20-minute start time threshold were included, yielding a 90% POD level at a gas flow rate of 2400 g/h (left). Detection reports meeting the 30-minute start time threshold were included (right). The 90% POD level is reached at a gas flow rate of 990 g/h.

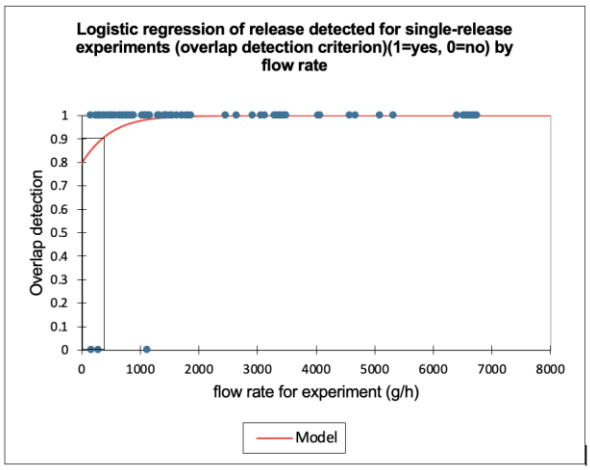


Figure 2. Probability of detection as a function of leak flow rate for experiments consisting of a single release. Detection reports that qualified as true positives based on the overlap criterion were included. For this data set, the 90% POD level is reached at a gas flow rate of 330 g/h.

For the set of all experiments (single-release and multiple-release experiments), and using the maximum release rates as the predictor parameter, the POD is greater than 90% for the lowest flow rates and then decreases as the flow rate increases (Figure 3). This is counterintuitive but reflects the fact that several of the multiple-release experiments with the highest release rates were not detected. These were events with relatively short gas release duration and/or unfavorable wind conditions.

Earthview Corporation
908 Dragon St
Dallas TX, 75207

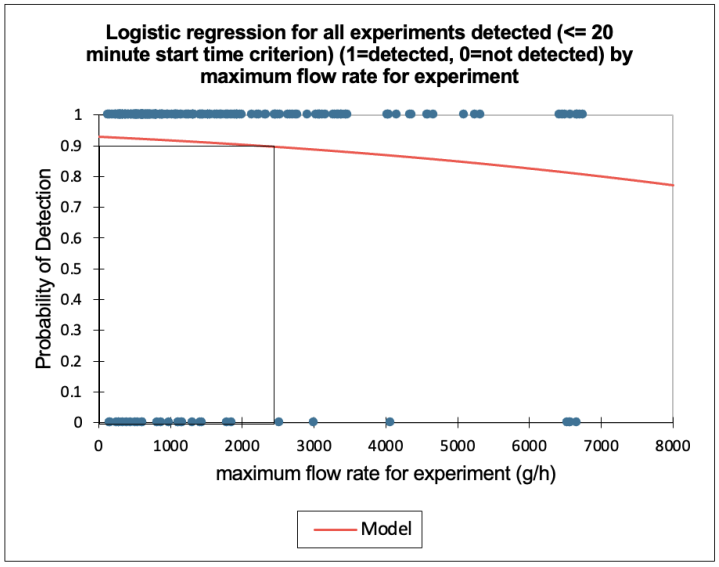


Figure 3. POD as a function of leak flow rate for all experiments, with the maximum release rate per experiment as the predictor parameter. Detection reports meeting the 20-minute start time threshold were included.

Since there is no physical reason why detection would decrease with increasing release rate, the behavior seen in Figure 3 likely indicates that flow rate is not the dominant parameter affecting the BluBird system's detection rates for this ADED test period. In fact, if we calculate POD as a function of release duration rather than release rate, we see the more logical pattern of increased POD with longer release times (Figure 4). For the 20-minute detection criterion, 90% POD is reached when the leak duration exceeds 2.2 hours, with the POD continuing to increase past that point. This is the behavior one would expect from point sensors - the longer a leak persists, the more likely it is that winds will transport leaking gas to a sensor. Using the 30-minute start criterion (Figure 6), the 90% POD occurs for leak duration of 1.7 hours.

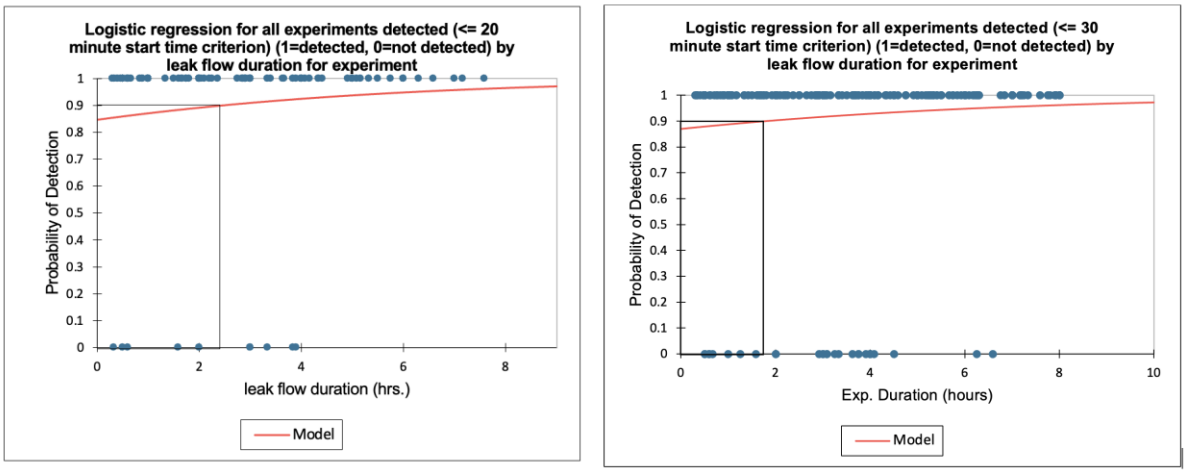


Figure 4. (Left) POD as a function of leak duration. Detection reports meeting the 20-minute start time threshold were included. (Right) POD as a function of leak duration for all experiments using detection reports meeting the 30-minute start time threshold.

Earthview Corporation
908 Dragon St
Dallas TX, 75207

Assuming POD is affected both by release rate and by leak duration, one way to capture this is by calculating the total gas release during experiments (e.g., release rate x release duration). This relationship is depicted in Figure 5. This POD plot suggests that a leak yielding at least 3800 grams should be detectable at the 90% level. In other words, at a leak rate of 400 g/h, the leak might be expected to be detected after about 10 hours' duration. As with all the results presented here though, this should be interpreted within the context of the METEC ADED test set-up and protocol.

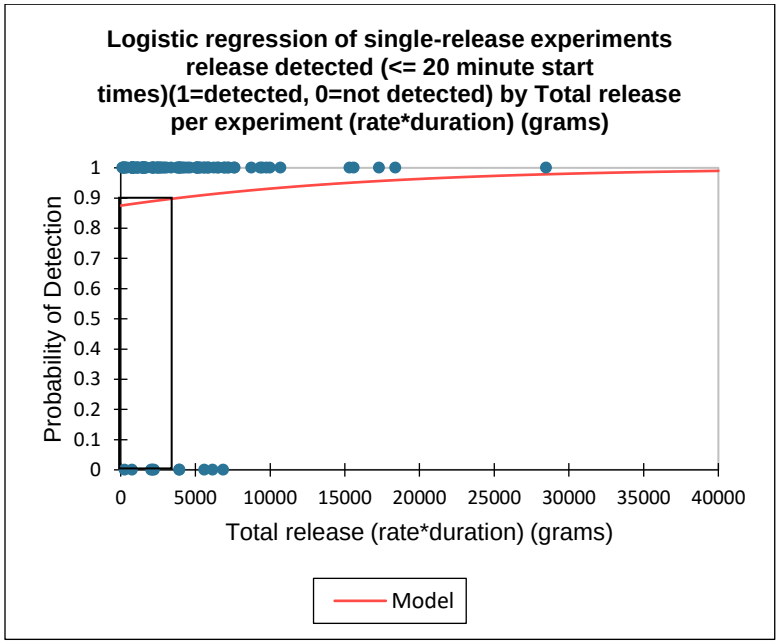


Figure 5. Probability of detection as a function of total natural gas released during experiments. Detection reports meeting the 20-minute start time threshold were included.

Notes of Caution Regarding the METEC ADED Testing Protocol and Analysis Methods

Although the current METEC ADED protocol is intended to be technology-neutral (Ilonze et al., 2024), the protocol and METEC's analysis methods include several aspects that arguably make the testing program more favorable for leak detection technologies that rely on imaging or scanning methods as compared to point sensors. METEC is working on a revised protocol for the future. However, to allow fair assessment of the different technologies, the test and analysis tradeoffs need to be acknowledged by METEC and their researchers and accounted for to the extent possible. We review some of these issues below. We understand the need to repeat the same methods from year to year so that results are comparable across the multiple ADED testing periods, but this should not rule out adding new analyses designed to address concerns, nor should it justify continuing to use methods (such as the curve fitting approach discussed below regarding probability of detection assessment) that may not be particularly appropriate.

Earthview Corporation
908 Dragon St
Dallas TX, 75207

Implications on Point Source Sensor Assessment of ADED's Focus on Multiple, Simultaneous Gas Releases: The most important factor in this regard is METEC's use of multiple, simultaneous gas releases from any of the 5 equipment groups on the test pad, with as many as 5 releases per experiment group. Failure of a technology to see all the simultaneous emissions during an experiment is treated as equivalent to failure to detect that any leak occurred at all. For example, consider two experiments: one experiment consists of a single release and the second experiment consists of 5 releases. Technology A detects the single release during the first experiment and 2 of the 5 releases during the second experiment. METEC's grading system that has been used to date assigns 3 true positives and 3 false negatives to Technology A. In contrast, Technology B misses the first experiment entirely but detects all 5 of the releases during the second experiment. This results in 5 true positives and 1 false negative. If the goal of the testing is to rate how many total leaks a technology detects, then Technology B ranks highest. This is the rating approach that METEC uses. But if one is interested in how often a technology detects presence of a leak on a pad, then Technology A would rank highest. METEC does not include this type of assessment.

At least some operators might think that is more important to know if there is a leak on their site rather than to know how many leaks there are at any given time. For one thing, an LDAR team dispatched to a site based on detection of a site-level leak is likely to find any other leaks that are present. But if the LDAR team is not dispatched at all, because the monitoring technology saw no leak on a site, then the leak will continue.

In the above example, a point sensor system is likely to perform more like Technology A while a scanning/imaging system is likely to perform more like Technology B. The main problem with METEC's approach to date is not that they use simultaneous leaks (with 81% of experiments in 2024 having more than one leak), it is that, in their publications (with the exception of a brief mention in Bell et al. [2020]), they do not point out the implications of their grading approach for different types of technologies. A simple and adequate solution would be for METEC to include, in their future publications and along with their current total-leak-detection approach, analysis that would compare technologies based on how many individual (and/or single-release) experiments were detected. This would show how good a particular technology is at alerting to site-level leaks. Based on published ADED results to date, there appear to be some significant differences in how point-source continuous monitors and imaging/scanning systems perform when rated using these two assessment methods.

Other METEC ADED Factors that can Negatively Affect Point Source Sensor Performance: In addition to METEC's grading method, some additional factors of the ADED protocol tend to be disadvantageous for point sensors. For example, the experiment durations are relatively short, averaging about 3 hours. Point sensors require variations in wind direction to carry gas from a leak source to a sensor, so short durations mean less time for one or more sensors to see a gas release. Also, there is some potential for residual natural gas to linger on the site (or be recirculated back onto the site by winds) in between experiments. This could potentially have a greater effect on point sensors than on scanning/imaging systems since in the METEC ADED protocol, detection reports are rejected if technology reports an event start time that is too

Earthview Corporation
908 Dragon St
Dallas TX, 75207

early. Residual gas can result in point sensor systems' attaching an earlier than actual start time to leak reports. As far as we are aware, METEC has not used any reference-grade instruments to verify background methane concentration in between experiments as part of the ADED test protocol.

METEC's Approach to Estimating 90% Probability of Detection (POD): It is typical to want to characterize a monitoring technology's performance in terms of the leak rate that can be detected 90% of the time. Technology providers are judged (fairly or not) on what 90% POD they can claim. To determine a useful POD value, the data used must encompass variations in leak rates, leak durations and leak locations over a range of weather variations and under realistic field conditions. ADED meets this need.

However, the actual calculation of POD can be very sensitive to the statistical method used, the range of factors affecting detections, as well as to assumptions that affect the true positive/false negative assessment. In the text below, we consider how some of these issues affect test results and how that in turn affects the assessed performance of technologies (specifically, Earthview's BluBird system). In particular, we review how analysis methodology can have a very large effect on estimated POD.

below, we show that using a more appropriate curve fitting method (logistic regression) along with slightly different criteria for leak detection classification (allowing as little as an extra 30 seconds for leak start-time thresholding) have a large effect on the estimated POD. We also show that, consistent with the ability of the BluBird system to detect small leaks, leak duration is more highly correlated with POD than is flow rate.

1. A Binary Classification Function Should be Used for POD Estimation

In their analysis of ADED results for 2020, Bell et al. (2020) used a binary logistic regression method to estimate POD. This method is well suited for predicting the likelihood of correct classification, such as ADED's "true positive" (TP) versus "false negative" (FN) classifications. However, Ilonze et al. (2024) instead chose to fit curves to the fraction of true positives divided by the sum of true positives and false negatives, as calculated over different bins of leak-rate ranges. A power law function was used for the fit, with the curve's intercept set to zero. The reason given for this switch is that the logistic regression model in some cases produced curves that were considered unrealistic since they yielded non-zero POD at zero emission rates. One can argue that this is not a good enough reason to abandon the strengths of the binary regression approach for a task like POD calculation (e.g., Binary Logistic Regression, 2024). In fact, METEC in their draft "METEC 2.0" protocol document specifically states that the TP/TP+FN ratio should not be used for POD estimation, and that a generalized linear model (such as binary logistic regression) be used[4].

The figures below illustrate the significance of this POD methodology as applied to Earthview's ADED 2024 results. Unlike the data sets discussed in the earlier sections, this data set includes all releases for all experiments (e.g., including multiple-release experiments). The classified

Earthview Corporation
908 Dragon St
Dallas TX, 75207

data consist of "1" where the release was assigned a TP and a "0" when an FN was assigned. The predictor variable is leak flow rate (except for Figure 8, where leak duration and total gas emitted per leak event are used).

Figure 6 compares results using logistic regression versus the power curve plot from Cheptonui et al. (2024). The results are visually similar but the results differ dramatically. in terms of estimated 90% POD. The logistic regression line intercepts the 90% POD axis at a leak flow rate of 9 kg/h whereas the power curve intersects at 76.5 kg/h. This involves projecting the line well beyond the range of test conditions, but was nevertheless applied this way in Cheptonui et al. (2024). As noted above, for applications like this where to goal is to predict the likelihood of a correct classification, a logistic function is more appropriate than is fitting a curve to the ratios of true positive/(true positive + false negative) calculated over release-rate bins.

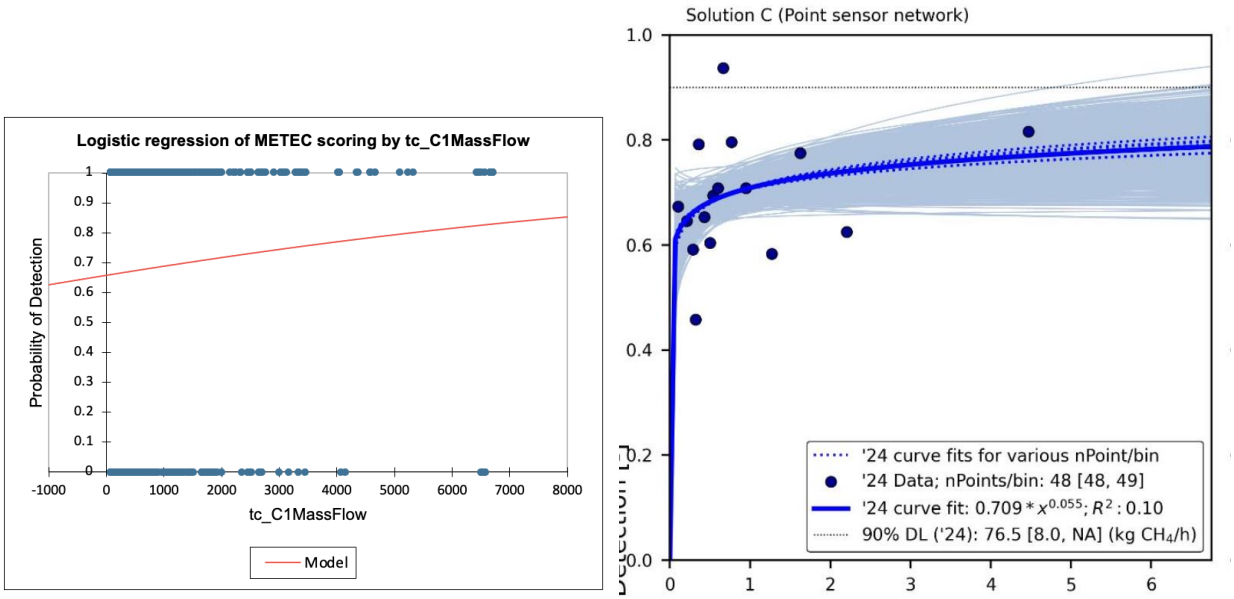


Figure 6. Logistic regression fit using binary classified detections (left) versus METEC's calculated power curve fit using ratios of true positives/(true positives + false negatives) calculated over release-rate bins (right).

2. Earthview's POD is More Highly Correlated to Leak Duration and Total Leak Amount than to Leak Rate

Regression fits to leak duration and total gas emitted per leak are given in Figure 7. Both of these fits have a higher level of significance (based on Chi-Square values) compared to the fit versus flow rate. (Note also the low R^2 of 0.10 in METEC's POD plot in Figure 6.) In terms of total gas emitted per leak, the estimated 90% POD for total gas release per leak is 14 kg. This could be a useful parameter to consider since in combines flow rate and duration. The

Earthview Corporation
908 Dragon St
Dallas TX, 75207

correlation suggests that it would be useful to consider both duration and rate to define POD for the BluBird system (and probably for other point sensors as well).

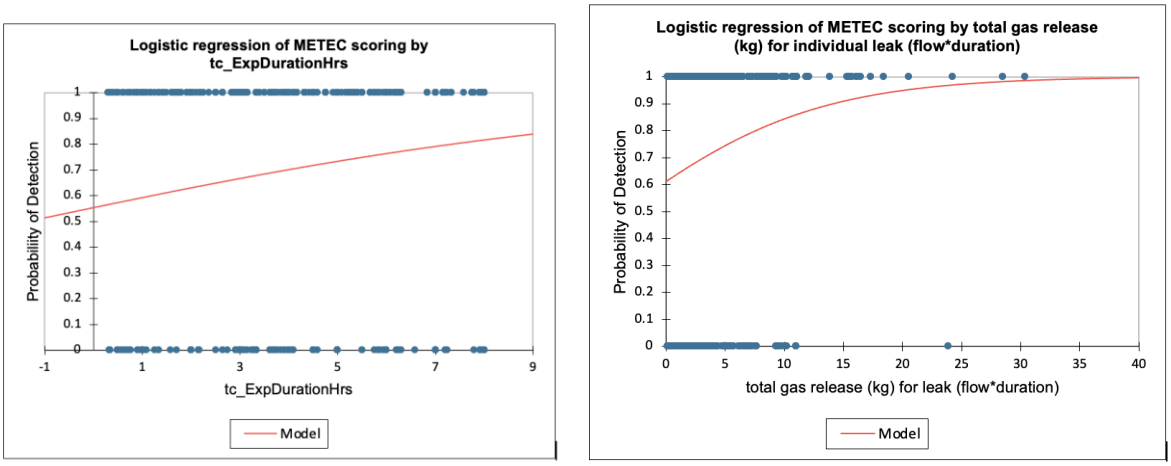
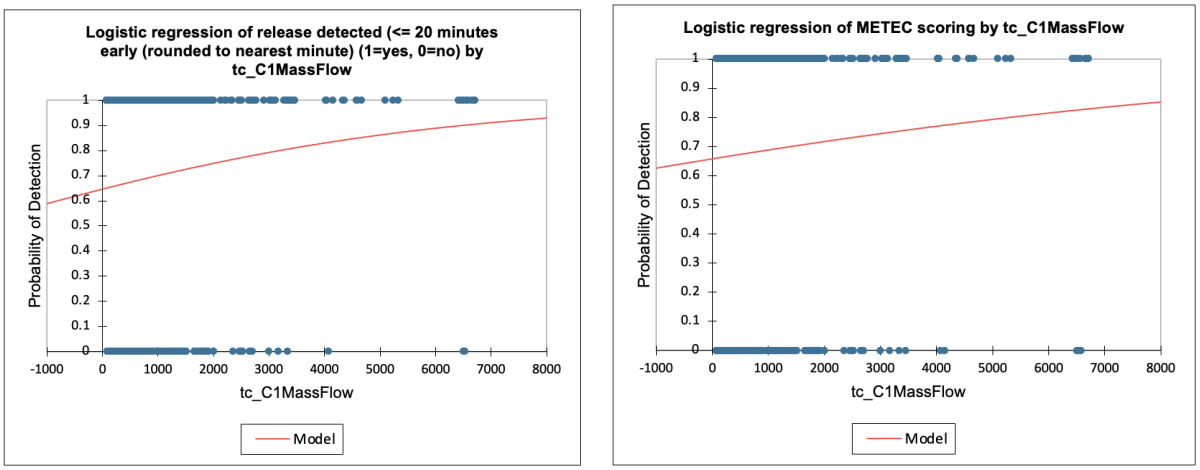


Figure 7. Logistic regression fits detection (binary; 0, 1) versus leak duration (left) and total gas released per leak (flow rate x duration) on the right.

3. The POD Estimate is Very Sensitive to the Inclusion of a Few Additional Detections

Slight differences in METEC's TP vs. false negative ("FN") grading protocol has large effects on estimated 90% POD. Figure 8 compares the METEC classification results using a 20.0 minute early start time threshold with using the start time rounded to the nearest minute. In other words, it treats start times within 30 seconds of the 20-minute threshold as a true detection. This slight change results in including an additional 6 detections as TPs (527 TPs vs. 533). The resulting 90% POD improves from 9000 g/h to 6400 g/h. If the start time threshold is relaxed to 30 minutes instead of 20 minutes, an additional 12 detections qualify as TPs, resulting in a 90% POD of 4800 kg/h. Allowing the extra 10 minutes time therefore cuts the 90% POD nearly in half in terms of release rates. Using the overlap classification criterion discussed earlier increases the true positive detections to 560 results in a 90% POD of 3700 kg/h.



Earthview Corporation
908 Dragon St
Dallas TX, 75207

Figure 8. Comparison of logistic regressions calculated using 533 TPs (left) and 527 TPs (right), where the 6 additional TPs result from allowing start times to be an additional 30 seconds earlier than the METEC protocol's start-time threshold of 20.00 minutes.

[1.] A successful detection is defined as an Earthview-issued report corresponding to a documented METEC leak (either a single leak per experiment or simultaneous multiple leaks per experiment). A METEC experiment consists of between one and five simultaneous releases from different locations on the test facility. Detection of an experiment therefore represents a detection of the presence of a leak at the site level.

[2]. In keeping with METEC's grading policy, a successful detection is required to have a reported start time no earlier than 20 minutes. Here, times are rounded to the nearest minute. For comparison, detections were also calculated using a start-time gap of 30 minutes. This takes into account that Earthview's automated processing system uses a 10-minute moving window to search for the start of a leak. If a leak is determined to have started during this window, the start time of the 10-minute window is assigned as the start time for the release. This could result in the reported start time being as much as 10 minutes earlier than the actual start time.

[3.] As discussed further below, there is also some indication that residual natural gas may have lingered on the site in between scheduled test releases. The BluBird system is sensitive to this, which may have resulted in early estimates of start times. With this possibility in mind, a separate criterion that defines a successful detection report as one overlaps substantially with the METEC release period was also tested. These overlap-based detections can have an earlier start time than 30 minutes but still result in distinct matches with METEC release periods.

[4]. *The following text is taken verbatim from the METEC 2.0 draft protocol plan (METEC, 2024):*

"Probability of Detection (POD)
The probability of detection (POD) evaluates the ability of the Solution to detect emissions and alert a user of the Solution that a response may be required. POD is always expressed as a curve (one independent variable) or surface (multiple independent variables) as a curve fit with detections (e.g. TP/FN) as binary values on the dependent axis. The POD curve should:

- Use clearly defined independent variables produced by the Test Center.
- Calculate the curve utilizing a *generalized linear model* or similar curve fitting approach which will accept binary dependent data as inputs.

The link function for the curve fitting should be asymptotic at 100% detection. A logistic function is preferred, but other asymptotic forms are also suitable. The Test Center will select the link function to best characterize the results.
Note: The ratio between TP and TP+FN reports *is not* a valid description of POD and should not be reported.
Reason: The mix of Controlled Release rates, durations, environmental conditions, and other factors are not necessarily characteristic of any Field Deployment, and this metric does not express the variation in POD over those variables."

References

Bell, C., T. Vaughn and D. Zimmerle, 2020. Evaluation of next generation emission measurement technologies under repeatable test protocols, Elem. Sci. Anthr., vol. 8, p. 32, doi: 10.1525/elementa.426.

Binary Logistic Regression, 2024. "Binary Logistic Regression", Science Direct literature survey, https://www.sciencedirect.com/topics/computer-science/binary-logistic-regression.

Cheptonui, F., E. Emerson, C. Ilonze, R. Day, E. Levin, D. Fleischmann, R. Brouwer, and D.J. Zimmerle, 2024. Assessing the performance of emerging and existing continuous monitoring solutions under a single-blind controlled testing protocol, ChemRxiv, working paper not yet peer-reviewed https://doi.org/10.26434/chemrxiv-2024-f1znb

Ilonze, C., E. Emerson, A. Duggan and D. Zimmerle, 2024. Assessing the progress of the performance of continuous monitoring solutions under a single-nlind controlled testing protocol, Env. Sci. Tech., vol. 58, issue 25, pp. 10881-11204, https://doi.org/10.1021/acs.est.3c08511

METEC, 2024. Controlled Test Protocol: Emission Detection and Quantification Protocol Revision 1.4 December 17, 2024 https://colostate-my.sharepoint.com/:w:/g/personal/c837377815_colostate_edu/EVJYH-4m7dpPnHYMBtOF2CIBEAxqSFWNv9Oddtu-g-pBug?e=XURkWM