Improving Consistency in Methane Emission Quantification from the Natural Gas Distribution System across Measurement Devices

Judith Tettenborn[1], Daniel Zavala-Araiza[1,2], Daan Stroeken[1], Hossein Maazallahi[1*], Carina van der Veen[1], Arjan Hensen[3], Ilona Velzeboer[3], Pim van den Bulk[3], Felix Vogel[4], Lawson Gillespie[4,5], Sebastien Ars[4], James France[6,7], David Lowry[6], Rebecca Fisher[6], and Thomas Röckmann[1]

[1]Institute for Marine and Atmospheric Research Utrecht (IMAU), Utrecht University, Utrecht, The Netherlands
[2]Environmental Defense Fund, Amsterdam, The Netherlands
[3]Netherlands Organisation for Applied Scientific Research (TNO), Utrecht, The Netherlands
[4]Climate Chemistry Measurements and Research, Climate Research Division, Environment and Climate Change Canada, Toronto, Canada
[5]Department of Physics, University of Toronto, Toronto, Canada
[6]Department of Earth Sciences, Centre of Climate, Ocean and Atmosphere, Royal Holloway, University of London, Egham, United Kingdom
[7]Environmental Defense Fund, London, United Kingdom
[*]Now at: Department of Renewable Energies and Environment, College of Interdisciplinary Science and Technologies, University of Tehran, Tehran, Iran

Correspondence: Thomas Röckmann (t.roeckmann@uu.nl)

Abstract.

Mobile real-time measurements of ambient methane provide a fast and effective method to identify and quantify methane leaks from local gas distribution systems in urban areas. The objectives of these methodologies are to i) identify leak locations for repair and ii) construct measurement-based emission rate estimates, which can improve emissions reporting and contribute to monitoring emission changes over time. Currently, the most common method for emission quantification uses the maximum methane enhancement detected while crossing a methane plume. However, the recorded maximum depends on instrument characteristics, such as measurement cell size, pump speed and measurement frequency. Consequently, the current approach can only be used by instruments with similar characteristics. We suggest that the integrated spatial peak area is a more suitable quantity that can eliminate the bias between different instruments. Based on controlled release experiments conducted with various different devices in four cities (London, Toronto, Rotterdam, and Utrecht), emission estimation methodologies were evaluated. Indeed, ~~the~~ when different analyzers were measuring in the same vehicle and from the same air inlet, the integrated spatial peak area was found to be a more robust metric across different methane gas analyzer devices than the maximum methane enhancement. A statistical function based on integrated spatial peak area is proposed for more consistent emission estimations when using different instruments. On top of this systematic relation between actual emission rate and recorded spatial peak area, large variations in methane spatial peak area were observed for the multiple transects across the same release point, in line with previous experiments. This variability is the main contributor of uncertainty in efforts to use mobile measurements to prioritize leak repair. We show that repeated transects can reduce this uncertainty and improve the categorization

into different leak categories. We recommend a minimum of three and an optimal range of 5-7 plume transects for effective emission quantification to prioritize repair actions.

1 Introduction

Mitigating methane (CH_4) emissions is an important step to combat climate change. The increase in atmospheric CH_4 has contributed approximately 30% of global warming since pre-industrial times (?)(?). CH_4 has an 82 times higher global warming potential over 20 years (?)(?), and a shorter atmospheric lifetime (9.1 $\pm$ 0.9 years (?))(?)) than CO_2. This offers the opportunity for reductions in climate warming by reducing CH_4 emissions, possibly avoiding temperature overshoots (?, ?)(??). General awareness of the importance for timely CH_4 mitigation is rising. More than 150 nations have joined the Global Methane Pledge, established during the UN Climate Change Convention Of Parties (COP 26) in 2021, pledging to reduce human-made CH_4 emissions in 2030 by 30% relative to 2020 (?). Following that(?). Following this, regulations specifically targeting CH_4 emissions have been adopted or are underway, for examplein , in Canada, Colombia, the EUand North America (?, ?, ?, ?), and the U.S. (????).

About a quarter of total global anthropogenic CH_4 emissions can be attributed to the oil and natural gas sector (?). The ? (?) . The ? suggests that deploying existing technological solutions could cut 75% of global oil and gas CH_4 emissions, at a cost less than 3% of the net income of the oil and gas industry in 2022. After the production segment, emission intensity is highest in the downstream distribution segment (?, ?). Regionally(??). Regionally, this sector can even present represent the largest share, especially in the EU, where the vast majority of oil and gas used is imported (?)(?).

Leak detection, quantification, and verification surveys are key to mitigate CH_4 emissions from the oil and gas supply chain. CH_4 analyzers on various plattforms platforms such as satellites, aircraft or vehicles have been utilized to detect and quantify fossil fuel related CH_4 emissions (?, ?, ?, ?, ?, ?, ?)(???????). Both rapid detection and reliable quantification can help to improve and prioritize capital-intensive leak repair efforts (?)(?). This is especially important since leak distributions have been found to be skewed, a few large leaks are being responsible for a major share of total emissions (?, ?, ?, ?, ?, ?)(??????) . Beyond the direct mitigation opportunity, reliable quantification can help to evaluate the scale of fugitive emissions and possible emission reductions over time, related to the Global Methane Pledge. Vehicle-based mobile measurements deploying fast CH_4 gas analyzers have proven to be an efficient method for quickly and effectively surveying distribution networks across urban areas. There is currently no universally accepted measurement and data analysis methodology in place. Most widely used is the statistical methodology developed by ? ? and later refined by ??. Based on controlled release experiment data (?)(?), they derived the following relation between the observed maximum excess CH_4 (in ppm) when crossing a plume and the release rate (r_E, in $\mathrm{L\,min^{-1}}$)

$$\ln([\mathrm{CH_4}]_{\max}) = 0.817 \cdot \ln(r_E) - 0.988. \quad (1)$$

In the original multivariate regression model suggested by ? ?, the release rate was used as the response variable, and maximum enhancement, integrated spatial peak area and an index of the plume kurtosis were used as predictors. ? ? argued

that maximum CH_4 enhancement proved to be the best predictor of the release rate and that additional predictors did not meaningfully improve emission rate estimates.

This transfer equation has been applied in several measurement campaigns in North America and Europe (?, ?, ?, ?, ?, ?, ?, ?, ?)(??????????). During each survey, hundreds to thousands of leak indications were detected. The emission estimates were also used to rank leak indications into different categories (high (>40 $\mathrm{Lmin^{-1}}$), medium (6-40 $\mathrm{Lmin^{-1}}$) and small (<6 $\mathrm{Lmin^{-1}}$) emissions), to assist with repair prioritisation decisions (?, ?)(??). However, several problems have been identified. Leak indications cannot always be confirmed through re-measurement, and gas providers cannot always identify the source of the emissions (?, ?, ?)(???). Furthermore, environmental factors, such as wind, but also dispersion and turbulence within the emission plume, lead to large uncertainties in measurements with corresponding over- and underestimations. Especially , within urban built environments, complex wind flow, re-circulation, and dispersion patterns play an important role (?, ?)(??). In addition, the approach developed in ? ? was derived for measurements using one specific device (Picarro G2301). By now, multiple CH_4 analyzers are in use, and it has been shown that CH_4 enhancements measured with different instruments can differ substantially (?, ?)(??). These discrepancies are expected, as different analyzers exhibit distinct instrument characteristics. Parameters such as cell volume, cell temperature, cell pressure, measurement frequency, and flow rate vary among instruments, all of which influence the detected shape of the CH_4 peak (?). Thus(?). Thus, the transfer function developed by ??, where estimated emission rates purely depend on the measured peak maximum, is not transferable to other analyzers. ?

? noted that the integrated spatial peak area is much more consistent between different instruments than the peak maximum . ? than the peak maximum between two different instruments operated in parallel during mobile measurement campaigns in Hamburg and Utrecht. ? conducted controlled release experiments and street-level measurements in Toronto, using two CH_4 devices. By applying Gaussian plume inversions to both CH_4 peak maximum and peak area, they concluded that emission rates should be estimated using either the peak area or, if relying on enhancement height, incorporating explicit modelling of the instrument response function to account for variations in cell residence time. ? found the maximum amplitude of CH_4 enhancement to be $50-80\%$ lower when sampled with a high inlet at 2.5 m compared to sampling close to the ground, whereas the integrated spatial peak area was better comparable between the measurements at different heights.

This study evaluates results from different controlled CH_4 release experiments in four cities. Predictors for statistical emission rate estimation are evaluated. Specifically, we evaluate the consistency between measured peak heights and spatial peak areas across eight different instruments when transecting the same emission plume with the same air inlet. We then propose an instrument independent instrument-independent transfer equation that uses the integrated spatial peak area. We evaluate how successful this transfer equation is in categorizing emissions into different categories , for single and multiple passes.

2 Methods

In this section, we first describe the controlled release experiments conducted to generate the dataset (Sect. 2.1). We then outline the data processing steps, including peak identification and calculation of the Spatial Peak Area (Sect. 2.2). Next, we present the approach for estimating emission rates from the measurements (Sect. 2.3). Finally, we evaluate the performance of

the quantification method (Sect. 2.4), considering both its categorization ability and the impact of the number of transects on estimation accuracy.

2.1 Controlled Release Experiments

Controlled release experiments (CRE) were conducted with a total of nine different analyzers in four different cities (London, Rotterdam, Toronto and Utrecht) situated in three different countries by different research groups (Tab. 1). The release locations are visualized in the ~~SI, Figure~~ Supplement Information (SI), Fig. S1 and an overview of the specific release rates per release location is given in ~~.~~ the SI, Tab. S2.

Table 1. Controlled release experiments: Overview of cities, release rates, inlet height and location. In all experiments, the release height was set at ground level.

height~~CRE (~~City & Experiment ~~)~~	Date	GHG An
London, UK, I Day-1	2019-09-10	G2301-m
London, UK, I Day-2	2019-09-11	G2301-m
London, UK, I Day-3	2019-09-13	G2301-m
London, UK, II Day-1	2024-05-13	LI-7810[c]
London, UK, II Day-2	2024-05-14	LI-7810[c]
Rotterdam, NL	2022-09-06	G2301[a], Mira Ult
Toronto, CA, Day-1	2021-10-20	G2401[a],
Toronto, CA, Day-2	2021-10-24	G2401[a]
Utrecht, NL, I	2022-11-25	G2301[a],
Utrecht, NL, II	2024-06-11	G2301[a],

[a] Picarro INC, Santa Clara, USA. [b] Los Gatos Research, San Jose, USA. [c] LI-COR Environmental, Lincoln, USA. [d] Aerodyne Research, Billerica, USA. [e] Aeris Technologies, Eden Landing Road Hayward, CA. [f] MIRO Analytical AG, Wallisellen, CH.

The controlled release experiment in Rotterdam was conducted on September 6, 2022. The location was selected to reflect common urban characteristics with houses, parked cars and overhanging trees (see SI, ~~Figure S1a~~Fig. S2). Methane (purity $> 99.9\%$) was released through a 1/4' O.D. teflon tube from two cylinders placed at a total of three locations along two connected streets at a wide range of flow rates (0.15~~–120~~ –120 $\mathrm{L min^{-1}}$). However, at two of the three locations, the rotameter used to adjust the release rate was suspected to not work properly. Therefore, only releases at location 1 (flowrate controlled electronically by an Alicat MCP-100SLPM, ~~5 – 120~~ 5–120 $\mathrm{L min^{-1}}$) are considered. The release occurred 1–3 m from the

street, with the inlet line moved to 5 m for flow rates above 40 $\mathrm{L\,min^{-1}}$. Atmospheric CH_4 mole fraction was measured while driving along the release locations, using five different instruments distributed over two vehicles. The first vehicle is Utrecht University's Air Quality car (UUAQ, Institute for Risk Assessment Sciences, Utrecht University), an Opel Astra with an air inlet on its roof (inlet height ca. 1.7 m, see SI, ~~Figure S2a~~Fig. S2). This car contained two cavity ring-down spectroscopy (CRDS) analyzers, model G2301 and G4302 ~~(Picarro INC, USA)~~ and a mid-infrared laser absorption spectroscopy analyzer MIRA Ultra~~(Aeris Technologies, CA)~~. Two instruments, a MGA10 analyzer ~~(MIRO Analytical AG, CH)~~ and TILDAS Dual Laser Trace Gas Analyzer~~(Aerodyne Research, USA),~~, were utilized in the measurement trailer of a semi-trailer truck operated by ~~TNO~~Netherlands Organisation for Applied Scientific Research (TNO). The inlet is on the side of the trailer around 2.5 m above ground level. In the morning, both vehicles drove separately. During the afternoon session, the G4302 and Mira Ultra analyzer were transferred to the TNO ~~semi-mobile~~ truck to facilitate better comparison between the measuring devices and the UUAQ vehicle ceased its mobile measurements.

Two experiments in Utrecht were conducted at the Utrecht Science Park with multi-storey office and service buildings to the sides of the street (see SI, ~~Figure S1b~~Fig. S3). On November 25, 2022, CH_4 was released simultaneously at two different locations. Two manual flowmeters (Krohne DK800/PV (25–250 NL/h) at location 1 and Krohne DK800/PV (500-5000 NL/h) at location 2) were used to measure the release rates (~~three different release rates~~ spanning from 2.18~~–15~~ –15 $\mathrm{L\,min^{-1}}$), which were controlled by the pressure reducer of the cylinder. CH_4 mole fractions were measured by the G2301 and G4302 devices, the same devices used during the Rotterdam campaign, on board the UUAQ car. The car was driving in a circle around two buildings, passing each emission point once per circle. On June 11, 2024 CH_4 was released simultaneously at ~~in total three different locations~~ the same site (0.15~~–100~~ –100 $\mathrm{L\,min^{-1}}$) and measured by the G2301 and Mira Ultra instrument installed in the ~~IMAU van (?) .~~ van of the Institute for Marine and Atmospheric Research Utrecht (?). Initially, the two release locations from the previous experiment were used. Due to power issues at location 1, the release point was moved across the street soon after the start. An Alicat device (high rates) and an MKS (PR 4000) controller (low rates) were used. Initially, the Alicat was at location 1 and the MKS at location 2. Midway, they were switched to enable all release rates at both sites.

Nearby London (on an open airfield near Bedford), two measurement campaigns were carried out, the first on September 10-13, 2019 ~~and~~ (35 and 70 $\mathrm{L\,min^{-1}}$) and the second on May 13 and 14, ~~2024.~~ 2024 (0.2–70.5 $\mathrm{L\,min^{-1}}$). A G2301 analyzer was used on all days in the first campaign, on September 10 additionally an Ultraportable Methane:Ethane Analyser ~~(uMEA, Los Gatos Research, San Jose, CA))~~ and on September 11 a LI-7810 $CH_4/CO_2/H_2O$ Trace Gas Analyzer ~~(LI-COR Environmental, USA)~~ were used. For the second campaign only data collected by the LI-7810 instrument were evaluated for this study. The driving pattern consisted of ~~multiple~~ parallel legs perpendicular to the estimated wind direction, gradually moving away from the source.

In Toronto, two CREs were carried out. In the Toronto industrial port lands neighbourhood, both a mobile bicycle-trailer-based laboratory equipped with a UGGA analyzer (inlet at 1.6 m above ground), and a vehicle based setup measuring with a G2401 analyzer (~~Picarro INC, USA,~~ inlet at 2.5m above the ground) were deployed on October 20, ~~2021.~~ 2021 (2.5–20 $\mathrm{L\,min^{-1}}$). The second experiment was carried out on October 24, 2021 on a parking lot, deploying the same vehicle-based set-up (0.1–5 $\mathrm{L\,min^{-1}}$).

The average driving speed during plume transects was 3.8±1.0 ms^{-1}, 5.6±0.9 ms^{-1}, 5.9±0.9 ms^{-1}, 3.0±0.5 to 3.9±0.4 ms^{-1}, 6.0±1.5 to 7.3±1.4 ms^{-1} and 4.0±0.4 to 7.5±0.9 ms^{-1} in Rotterdam, Utrecht I, Utrecht II, London I, London II and Toronto respectively. The median distance between the location where the plumes were detected and release location was 20 m, 20 m, 21 m, 20 to 25 m, 21 to 22 m and 17 to 24 m, which is typical for urban gas distribution networks (see SI, Sect. S6). More detailed descriptions of the measurement devices~~and~~ , experimental set-up and the CH_4 timeseries can be found in the Supplementary Information ~~, ?, ? and ?~~(SI, Sect. S1–S4), **?**, **?** and **?**.

2.2 Data Treatment, Peak Identification, Determining Peak Maximum and Spatial Peak Area

The raw measurements were calibrated and corrected for inlet delay and a delay between different instruments (see SI, Sect. S3). A centred 5 minute moving window was applied to determine the atmospheric CH_4 background level at each point in time. The background level was defined as the 10^{th} percentile of the CH_4 mole fraction measurements, which was assessed to represent the background well (comparison is given in SI, Sect. S5). A peak was identified as a CH_4 enhancement reaching above 102% of background level. CH_4 enhancements in the calibrated CH_4 dataset were detected utilizing the python *scipy* function *find_peaks* ~~(?)~~(**?**). Those individual peaks were then manually quality-checked for overlap of peaks, flawless function of all instruments deployed, validity of the transect and car speed. When multiple instruments recorded measurements in one vehicle, the peak finding algorithm was applied to only one of the instruments. With a manual quality check it was ensured that the peak was valid for all instruments. For the London ~~dataset~~ datasets, peaks obtained at a distance further than 75 m from the source were omitted to keep the distance within the same limits as for the other CREs. The maximum CH_4 enhancement within the time interval of each peak was determined for each instrument. The peak finder algorithm was applied to the G4302 device for the measurements on the ~~UU~~ UUAQ car in Rotterdam and in Utrecht I, the G2301 instrument for Utrecht II, the MGA10 device for the measurement on the TNO truck in Rotterdam, the uMEA and G2301-m device during the CRE in London I for Day 1 and Day 2 respectively. The peak area was integrated over space rather than time to take different driving speeds into account. To convert the time series to the spatial coordinate, the CH_4 mole fraction enhancement at time t_{i+1} ($c(t_{i+1})$, in ppm, after subtraction of the CH_4 background level) was multiplied with the measurement time step of the individual instrument ($t_{i+1} - t_i = \Delta t$) and the velocity of the vehicle averaged over the whole peak duration ($\bar{v}_{\mathrm{peak}}$), yielding the integrated spatial peak area ($[\mathrm{CH_4}]_{\mathrm{area}}$) in $\mathrm{ppm} * \mathrm{m}$.

$$[\mathrm{CH_4}]_{\mathrm{area}} = \sum_{i=0}^{n} \Delta t \cdot c(t_{i+1}) \cdot \bar{v}_{\mathrm{peak}} \tag{2}$$

It is important to note that this spatial peak area does not correspond to the integration of CH_4 enhancement of the physical CH_4 plume in the environment across a 2D plane in space. Rather it represents a linear 1D fraction of the plume, that is described by the driving track.

2.3 Emission Rate Estimation

An ordinary least-squares regression model (*scipy.stats.linregress* library, ~~?~~**?**) was applied to the spatial peak areas of the combined dataset. The natural logarithm (ln) of the known release rate was used as the independent variable and the ln of

integrated spatial peak area of CH_4 enhancements as the dependent variable.

$$\ln([CH_4]_{area}) = a_1 \ln(r_E) + a_0 \quad (3)$$

The fitting was performed using the entire dataset, without separating it into training and testing data. To assess the conformity of the data with assumptions underlying a linear regression (normality, linearity, independence and homoscedasticity), several analyses were carried out (SI, Sect. S7). A similar fit was applied to the maximum excess CH_4 ($\ln([CH_4]_{max}) = a_1^{max} \ln(r_E) + a_0^{max}$) for comparison to the equation from ??.

To infer emission rate estimations based on measurements, the linear regression needs to be solved for the release rate. Then the equation can be applied to measurements (the Weller eq. to the peak maximum measurements, the Area eq. to the corresponding area measurements, for both cases background levels subtracted from CH_4 measurements). Sometimes the algorithm produces estimates far outside the calibration range of the method, therefore a cap of 200 $\mathrm{L min^{-1}}$ was imposed on emission rate estimations for the following evaluation.

2.4 Evaluating Quantification Performance

2.4.1 Categorization

~~To stay consistent with previous studies, release rates were classified into four different categories (< 0.5 $\mathrm{L min^{-1}}$-Very low, $0.5 - 6$ $\mathrm{L min^{-1}}$-Low, $6 - 40$ $\mathrm{L min^{-1}}$-Medium and > 40 $\mathrm{L min^{-1}}$-High). For each leak indication~~ For each valid plume transect (i) in the dataset used to derive the regression model, emission rates were estimated utilizing both estimation methods, ~~using~~ applying the inverse of Eq. 1 and Eq. 3.

$$r^i_{\mathrm{E,Weller\ eq.}} = \exp\left(\frac{1}{a_1^W}\left(\ln([CH_4]_{max})^i - a_0^W\right)\right) = \exp\left(\frac{1}{0.817}\left(\ln([CH_4]_{max})^i + 0.988\right)\right) \quad (4)$$

$$r^i_{\mathrm{E,Area\ eq.}} = \exp\left(\frac{1}{a_1^A}\left(\ln([CH_4]_{area})^i - a_0^A\right)\right) \quad (5)$$

The superscript W (Weller eq.) and A (Area eq.) differentiate the regression parameters. Subsequently, an estimated category was assigned to each peak given these inferred emission rates. To remain consistent with previous studies, release rates were classified into four different categories (< 0.5 $\mathrm{L min^{-1}}$-Very low, $0.5 - 6$ $\mathrm{L min^{-1}}$-Low, $6 - 40$ $\mathrm{L min^{-1}}$-Medium and > 40 $\mathrm{L min^{-1}}$-High). This approach follows ??, but is extended by a category for very small emissions as used in ??. For each group of peaks belonging to the same category, the percentage of correctly classified peaks was calculated, along with the percentages of peaks that were erroneously categorized into other categories.

2.4.2 Percentage Difference in Emission Estimation as Function of Number of Transects

As will be shown below, variability in the plume shape causes large differences in observed peak maxima, spatial peak areas and thus derived emission rates for individual transects at the same actual emission rate. This variability can be reduced by evaluating the average of several transects at the same emission rate. Following the analysis in ~~?~~ ? the effect of number of detections per CH_4 source on variability in estimated emission rate was explored.

Two approaches were followed to evaluate the emission quantification:

1. Comparison against the mean emission rate calculated from the leak indications (following ??).
2. Comparison against the true release rate, which is known for our experiments.

To calculate the mean emission rates for the former comparison, the average natural logarithm of the integrated spatial peak area among all observed instances associated with each release rate j was computed.

$$\text{Mean } \ln([\mathrm{CH_4}]_{\text{area}})^j = \frac{1}{n}\sum_{i=1}^{n} \ln([\mathrm{CH_4}]_{\text{area}})^i. \tag{6}$$

For each release rate, one mean emission rate estimation $r^j_{\text{E,mean}}$ was obtained by applying the Area eq. to the calculated mean $\ln([\mathrm{CH_4}]_{\text{area}})$. Subsequently, a Monte Carlo simulation was performed where we randomly selected between 2 and 10 emission peaks i at each release rate, and averaged the natural logarithm of the integrated spatial peak areas. Measurements obtained by different instruments during the same transects were treated as separate peaks. The theoretical number of possible subsets ~~N~~ N from a certain size of set ~~M~~ M with replacement is $\binom{M+N-1}{N} = \frac{(M+N-1)!}{N!\cdot(M-1)!}$. This gives for example between 465 and $6 \cdot 10^7$ combinations for N between 2-10 and $M = 30$. This procedure was repeated 2000 times for each release rate and number of detections. That means for $N = 3$, three peaks were randomly sampled 2000 times from a given release rate, yielding 2000 emission rate estimations $r^{k,j,N}_{\text{E,sim}}$ ($k \in [1, 2000]$) for each release rate. For each of those 2000 emission rate estimations $r^{k,j,N}_{\text{E,sim}}$ the percentage difference to the mean emission rate estimation $r^j_{\text{E,mean}}$ and the known release rate $r^j_{\text{E,true}}$ were calculated:

$$\text{Percentage Deviation } \Delta\%^{k,j,N}_{\text{mean}} = \frac{r^{k,j,N}_{\text{E,sim}} - r^j_{\text{E,mean}}}{r^j_{\text{E,mean}}} \cdot 100\% \tag{7}$$

$$\text{Percentage Deviation } \Delta\%^{k,j,N}_{\text{true}} = \frac{r^{k,j,N}_{\text{E,sim}} - r^j_{\text{E,true}}}{r^j_{\text{E,true}}} \cdot 100\%. \tag{8}$$

Then, an average percentage deviation for each release rate j and number of transects N was determined for both cases.

$$\text{Mean Percentage Deviation } \overline{\Delta\%}^{j,N} = \frac{1}{2000}\sum_{k=1}^{2000} \Delta\%^{k,j,N}. \tag{9}$$

Finally, an average over the different release rates J for each number of transects $N \in [2, 10]$ was calculated for both cases.

$$\text{Overall Mean Per N}\text{Overall Mean Per } N = \frac{1}{J}\sum_{j=1}^{J} \overline{\Delta\%}^{j,N} \tag{10}$$

The analysis was done for each unique release rate and location pair. Experiments with fewer than 10 transects were filtered out, as the Monte Carlo analysis samples up to 10 measurements (~~N~~ N ranging from 2 to 10) from the available transects. This leaves 35 of the original 55 release rates. Lastly, another categorization analysis was conducted with the goal to investigate the classification performance when incorporating multiple transects. For each of the 2000 mean emission rates per release rate and number of transects, a category was assigned and evaluated.

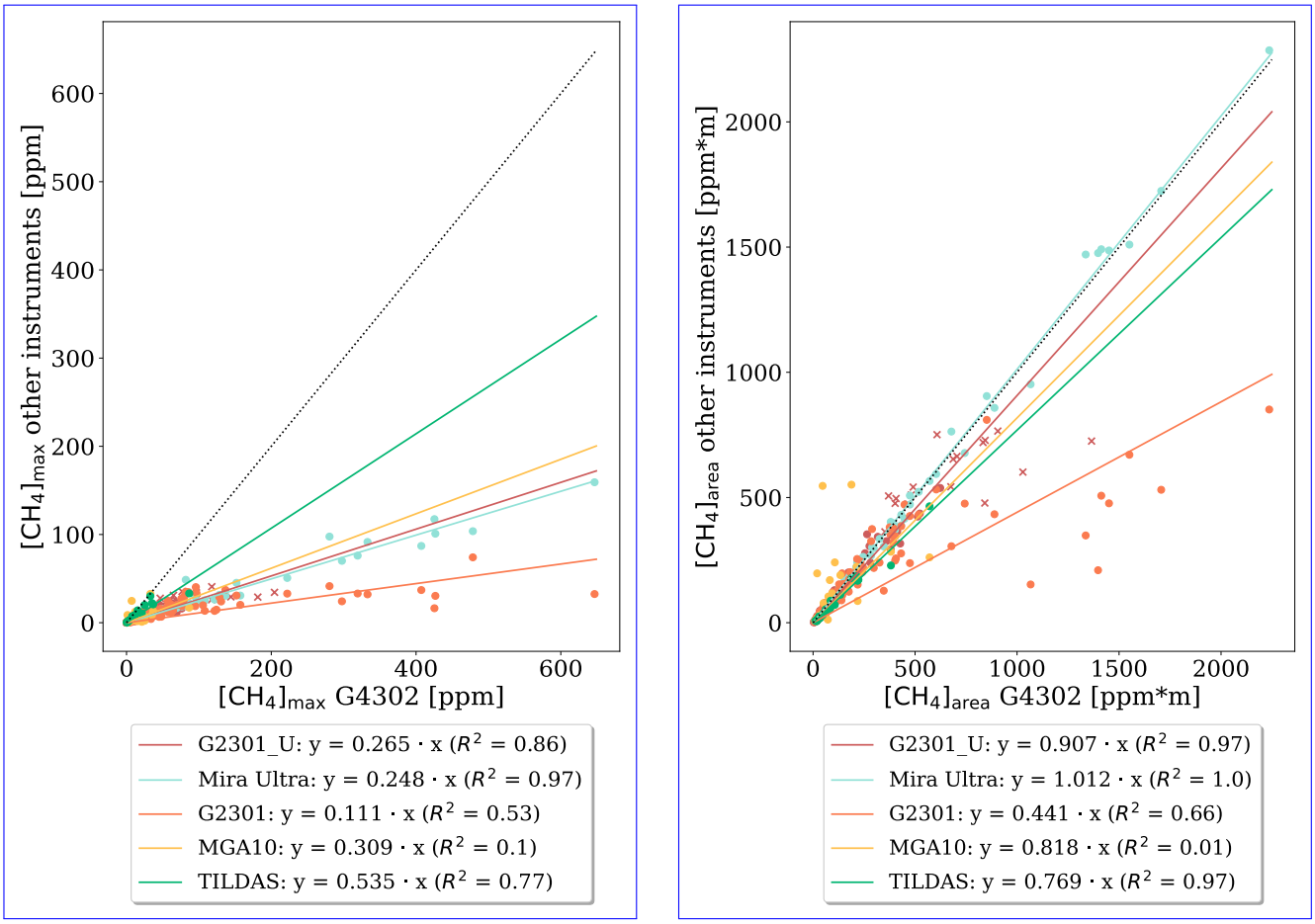
3 Results and Discussion

In this section, we present and analyze the key findings of our study. First, we evaluate instrument performance by comparing peak maximum and spatial peak area as metrics for emission quantification (Sect. 3.1). Next, we introduce a method for converting spatial peak areas into emission rate estimates (Sect. 3.2). We then explore how survey data can be used to support repair prioritization (Sect. 3.3). Additionally, we assess the benefits of multiple transects in improving emission estimates (Sect. 3.4). Finally, we propose a method for the routine implementation of leak surveys based on our findings (Sect. 3.5).

3.1 Instrument Performance: Peak Maximum and Spatial Peak Area

~~Peak Maximum Spatial Peak Area Comparison of peak maximum (a) and spatial peak area (b) from different instruments deployed in Rotterdam and Utrecht (subscript '_U'), shown are the data points and a linear regression fit with intercept 0 for each instrument. The results from the G2301, Mira Ultra, MGA10 and TILDAS devices are plotted on the y-axis and the results from the G4302 instrument on the x-axis. The black dotted line represents the 1:1 line. (For the G2301 analyzer, peaks exceeding a maximum of 20 ppm are marked with an 'x' and excluded from the fitting process.)~~

Fig. 1a shows that the comparison of measured peak maxima among different instruments reveals very strong systematic discrepancies. Compared to the G4302, all other instruments (Mira Ultra, G2301, MGA10, and TILDAS) strongly underestimate the peak maximum. Specifically, the maxima recorded by the TILDAS are only half ~~compared to those of~~ of those measured by the G4302, while the Mira Ultra, G2301, and MGA10 show peak maxima that typically are only between 10% to 30% of the G4302 readings. Fig. 1b shows that the evaluated spatial peak areas are much more consistent between instruments, with slopes between ~~0.63~~ 0.44 and 1.01. The coefficient of determination R^2 is notably higher for the area fit when compared to the maximum fit for ~~most~~ all instruments. R^2 for G2301_U, Mira Ultra, and TILDAS ~~all~~ exceed 0.96. The MGA10 device stands out as an exception, as it demonstrates poor R^2 values for both maximum and area fits. Both ~~,~~ measured maximum and area values are more scattered and exhibit strong deviations from the other instruments. Interestingly, the G2301 analyzer aligns better with the G4302 for data collected in Utrecht than in Rotterdam (indicated by a higher slope). This reflects the better alignment of instrument measurements for smaller peaks; in Rotterdam, very high release rates were included. A similar pattern arises for the evaluation of the CRE in London ~~(~~I (SI, Fig. S21). Both the uMEA and LI-7810 instrument measurements align more closely with G2301 measurements when assessing spatial peak area rather than peak maximum.

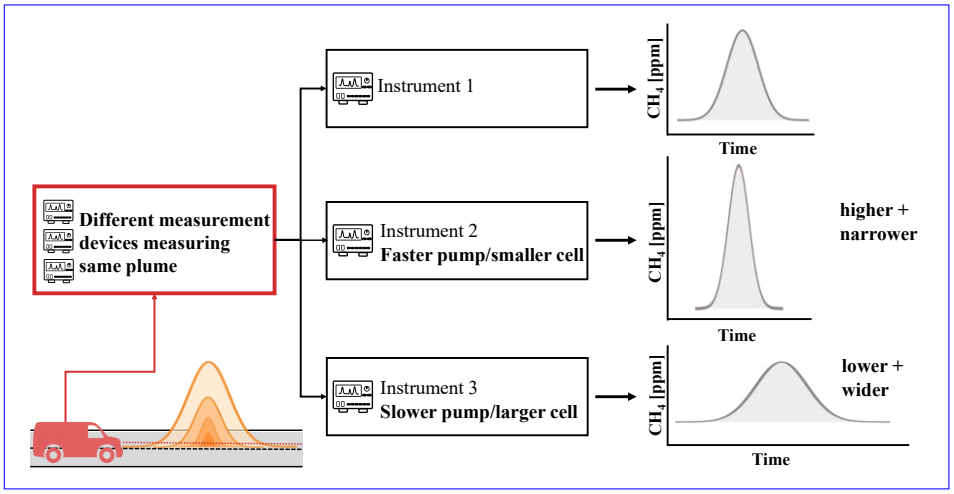


(a) ~~Flow Rate and Cell Volume~~Peak Maximum (b) ~~Measurement Frequency~~Spatial Peak Area

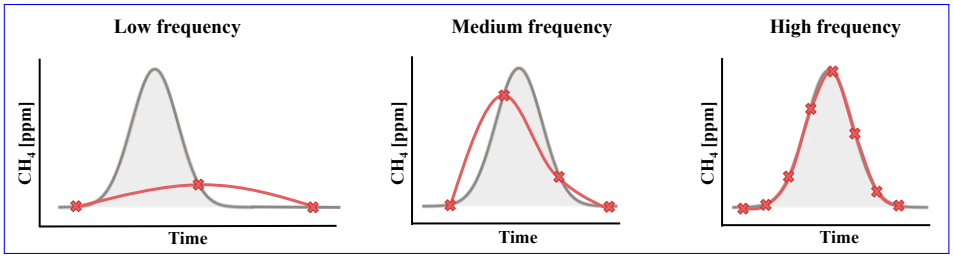
Figure 1. ~~Influence~~ Comparison of ~~the instrument characteristics flow rate, cell volume~~ peak maximum (a) and spatial peak area (b) from different instruments deployed in ~~dependence of cell temperature~~ Rotterdam and ~~pressure~~Utrecht I (subscript '_U'), shown are the data points and ~~measurement frequency~~ a linear regression fit with intercept 0 for each instrument. The results from the G2301, Mira Ultra, MGA10 and TILDAS devices are plotted on the ~~peak shape detected~~y-axis and the results from the G4302 instrument on the x-axis. The black dotted line represents the 1:1 line. (For the G2301 analyzer, peaks exceeding a maximum of 20 ppm are marked with an 'x' and excluded from the fitting process.)

Our analysis demonstrates that different CH_4 analyzers show much better consistency in the integrated CH_4 spatial peak area compared to the maximum CH_4 enhancement. This finding is supported by conceptual considerations: Measured mole fractions are influenced by instrument characteristics, mainly flow rate, cell volume, and measurement frequency (Fig. 2). A higher flow rate or smaller cell volume will generally make measured mole fraction profiles sharper and higher. Lower flow rates and larger cell volumes, on the other ~~side~~ hand, lead to smoothing effects, rendering the CH_4 peak smaller and wider ~~(? , ?)~~(??). While the peak maximum is strongly affected by ~~those~~ these characteristics, the integrated area is not. Each molecule

that enters an instrument is measured, either as part of a wide and low or a narrow and high peak. ~~? and ?~~ ? and ? made similar observations when comparing different analyzers. In this study, we generalize these results by demonstrating the same effect across eight different instruments, enhancing the robustness of this conclusion. Beyond differences in instrument performance, ~~? ?~~ noted the area to be more robust against differences in inlet height, further supporting the argument for using the integrated spatial peak area as a more reliable metric for quantifying CH_4 enhancements. In ~~the future , even more~~ future mobile surveys, several different instruments are expected to be deployed ~~across~~ to survey local natural gas distribution networks. It is therefore important to move beyond the maximum enhancement as emission estimation metric and use the spatial peak area instead.



(a) Flow Rate and Cell Volume

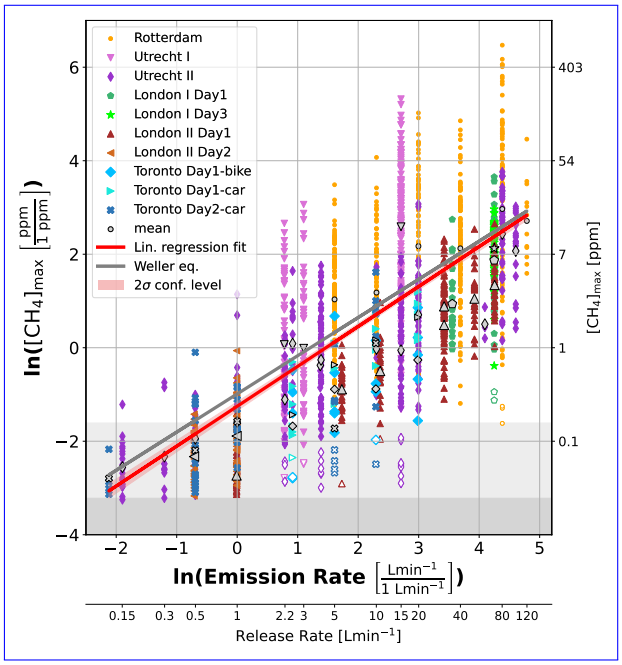


(b) Measurement Frequency

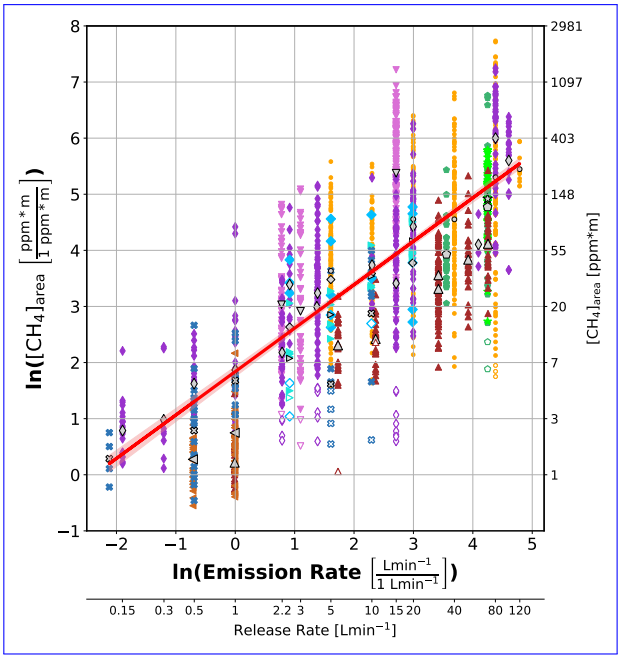
Figure 2. Influence of the instrument characteristics (a) flow rate, cell volume (in dependence of cell temperature and pressure) and (b) measurement frequency on the peak shape detected.

3.2 Converting Spatial Peak Areas to Emission Rates

Fig. 3b shows the recorded spatial peak areas as a function of the known release rates (double logarithmic axes) from all controlled release experiments. As anticipated, higher CH_4 release rates in general correspond to higher observed spatial peak area measurements. A linear fit can be applied to these data and the fit equation is proposed as an empirical model



(a) Peak Maximum



(b) Spatial Peak Area

Figure 3. Correlation of the natural logarithm of the peak maximum enhancement (a) and spatial peak area (b) with the natural logarithm of the release rates for all controlled release experiments reported in this manuscript (except London I Day 2). Different cities are indicated by different colours (Rotterdam-orange, Utrecht I-pink, Utrecht II-dark purple, Toronto-blue, London I-green and London II-brown). Black markers indicate mean values per release rate and city, unfilled markers indicate potential outliers. The second x-axis indicates release rates deployed, the red lines are linear regressions to all data. The Weller equation (grey line in (a)) is displayed ~~in gray~~ as a comparison ~~(a)~~ and the light (dark) gray area indicates peaks below 110% (102%) of background level.

relating integrated spatial peak area of CH_4 enhancements ($[CH_4]_{area}$, in $ppm * m$, background levels subtracted from CH_4 measurements) to emission rates (r_E, in $Lmin^{-1}$):

$$\ln([CH_4]_{area}) = 0.774 \cdot \ln(r_E) + 1.84. \quad (11)$$

In practice, emission rate estimations will be derived from area measurements. To achieve this, Eq. 11 can be solved for r_E.

$$r_E = \exp(1.292 \cdot \ln([CH_4]_{area}) - 2.377) \quad (12)$$

The natural logarithm of the spatial peak area associated with a leak indication can then be inserted in the equation. If multiple transects were taken, the average logarithm of the spatial peak area values $\overline{\ln([CH_4]_{area})}$ corresponding to the same leak indication should be used.

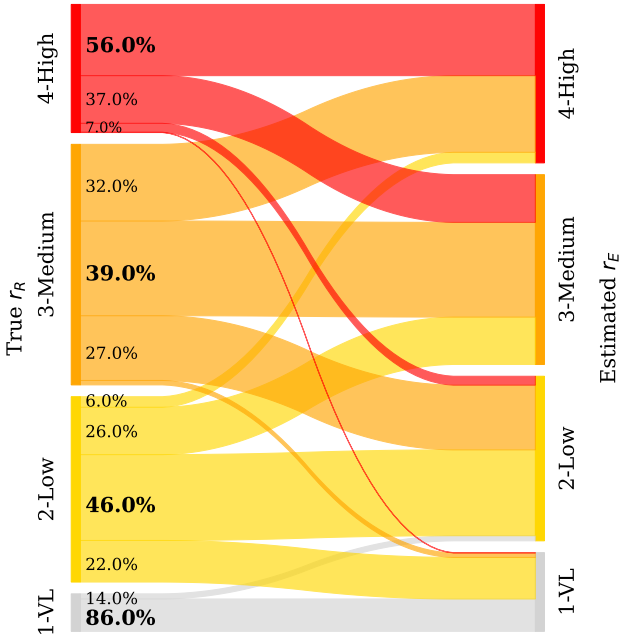
A similar linear regression fit was applied to peak maxima data: ~~$\ln([CH_4]_{max}) = 0.854 \cdot \ln(r_E) - 1.25.$~~ $\ln([CH_4]_{max}) = 0.854 \cdot \ln(r_E) -$ (Fig. 3a). The inferred regression equation for the maximum shows good agreement with the model proposed by ??.

3.3 Using Survey Data for Repair Prioritisation

A wide range of peak maxima and areas is observed for each release rate. This illustrates the shortcomings of the quantification of emissions for individual leak indications using this simple statistical method. This spread reflects the nature of turbulent dispersion of plumes. Environmental factors, in particular built environment and meteorology, e.g. convection, complex wind patterns within urban areas (back circulation, wind channelling, blockages), turbulence and diffusion, play a major role in determining the location, shape and CH_4 mole fraction of the meandering CH_4 plume ~~(?, ?, ?)~~(???). Still, a greater CH_4 release increases the likelihood of detecting higher CH_4 mole fractions. This is the relation that is captured in the proposed transfer equation. Yet, changing wind conditions or turbulence can still cause the main plume to become highly diluted or be transported away from the street, leading to smaller peak maxima and spatial peak areas. At the extremes, high CH_4 mole fractions are measured in one transect, and no peak at all in another one ~~(?).~~ (?). Ideally, measurements should be conducted downwind of the emission source and perpendicular to the wind. This approach is commonly used when targeting individual emitters, such as oil and gas processing stations or farms (??). However, when measuring emissions from urban gas distribution networks, the built environment imposes constraints on positioning. Incorporating wind measurements from the survey vehicle could potentially improve emission estimates and should be explored in future research.

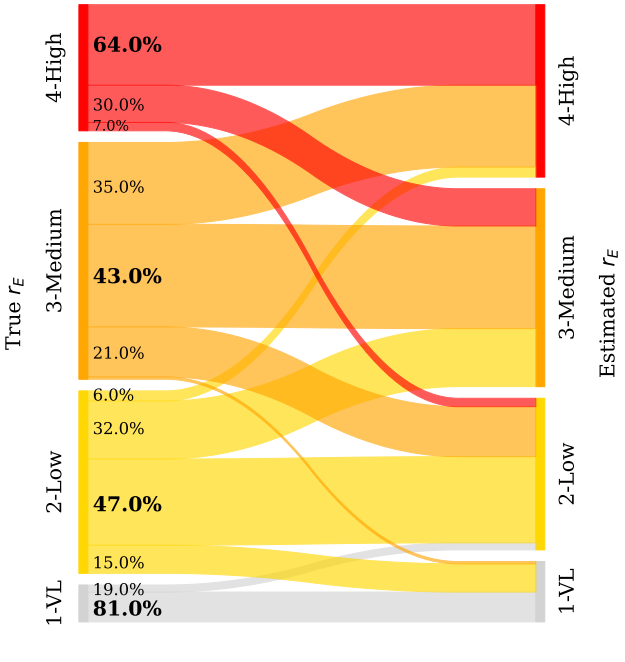
Additionally, some assumptions underlying the application of a linear regression are partially violated (homoscedasticity, normal distribution of errors). However, the premises were deemed sufficiently met to justify the use of linear regression, allowing for a straightforward and practical statistical inference model. However, as all experiments included in this study were executed during daytimes (measurements were conducted between 09:00-18:00 local time), the method might not be suitable for the evaluation of nighttime measurements. During the night, the atmospheric boundary layer becomes more stable, which suppresses updrafts and leads to an accumulation of CH_4 emissions at the surface. This causes higher measured mole fractions, increasing both the peak maximum and spatial peak area measurements; hence most likely considerable overestimation of the emission rate.

Similarly, though to a lesser extent, variations in solar irradiation during the day can influence updraft behavior. Around noon, stronger solar irradiation is expected to enhance updrafts, which could reduce ground-level methane concentrations. Only a few release rates were deployed twice on the same day, preventing a conclusive analysis of measurement comparisons at different times. We recommend that future experiments investigate diurnal effects more systematically.



(a) Weller eq., based on Peak Maximum

~~Area eq.~~



(b) ~~Categorization performance of the~~ Area ~~and Weller method (including data from all 6 CRE)~~eq.~~The left y axis represents the true emission rate r_E, where the width of the bars indicate the amount of plumes belonging to each emission category (categories: 1-Very Low, 2-Low, 3-Medium and 4-High). The right y axis represents the categories estimated by the statistical model~~

Deviation of emission rate estimation from true emission rate can be very high for an individual gas leak, especially for single transects. For our set of experiments under- and overestimation range from -100% to +2700%. To facilitate a pragmatic repair priorization, categorizing leaks can be an important tool. We quantified the emission rate of all leak indications and subsequently assigned them to one of the four categories described above. Fig. 4 illustrates the categorization performance based on quantification using the spatial peak area (a) and peak maximum (b). It displays the percentage of peaks correctly classified (bold number) and also shows in which categories the rest is falsely overestimated or underestimated. A suitable benchmark for categorization is a 25% accuracy rate, reflecting the expected success when peaks are randomly assigned to one of the four categories.

The majority of very low and high peak indications are correctly attributed. Peaks in the medium category are less well classified, with a large portion being ~~overestimated~~either over- or underestimated. Overall, the proportion of peaks that are misclassified by more than one category is very low. Around 7% (22%) of category 4 (3) peaks are underestimated into category 1 or 2, when applying the area eq. This means that the vast majority (78% to ~~93~~94%) of higher emission peaks from categories 3 and 4 are correctly identified as ~~high~~medium–high, effectively differentiating them from low emission peaks. At the same time, 38% (0%) of category 2 (1) emission rates are falsely identified as high emission rates. The same analysis using the Weller eq. yields comparable results, with a slightly lower categorization performance except for the lowest emission category (Fig. 4a).

It is important to note that these categorization performance rates vary across our set of release locations and are therefore not directly transferable to any leak populations in urban areas. The quantification and categorization performance differs for different locations, for example, the categorization precision for the London I data is 62% for category 4, but 88% for category 3, much higher than for the average (SI, ~~Figure~~ Fig. S22 and S23). Given the local circumstances regarding built environment and meteorological conditions, peak measurements can be systematically lower or higher compared to the mean of our dataset. However, the numbers can still be considered as an indication of typical categorization performance.

In ~~?~~ ?, a categorization success rate of over 80% for category 2 and 3 was reported, which is higher than for our dataset. For the highest category, their performance was lower (38%) than ours (64%). The differences might be due to the specific local conditions where their controlled release took place. In practice, $> 80\%$ of urban leaks have a low emission rate ~~(?, ?,~~ (??), category 1 was not assessed in those publications).

Overall, our results show that the categorization approach for leak repair prioritization would perform far better than making repair decision by chance. Therefore, this approach offers a potential to reduce emissions since larger leaks can be targeted first. The Area eq. does not consistently outperform the Weller model in terms of categorization performance, but it yields more consistent results across different instruments, which would facilitate comparison of measurement surveys performed with different instruments.

3.4 Benefit of Multiple Transects

~~Calculated mean emission rate~~ ~~True release rate~~ ~~Mean absolute percentage deviation of emission estimation from the (a) calculated mean emission rate and (b) true release rate as a function of the number of transects. Each line represents one~~

release rate at one location, and was created by a Monte-Carlo analysis with 2000 repetitions. The total mean over all release rates considered in this analysis is shown in black.

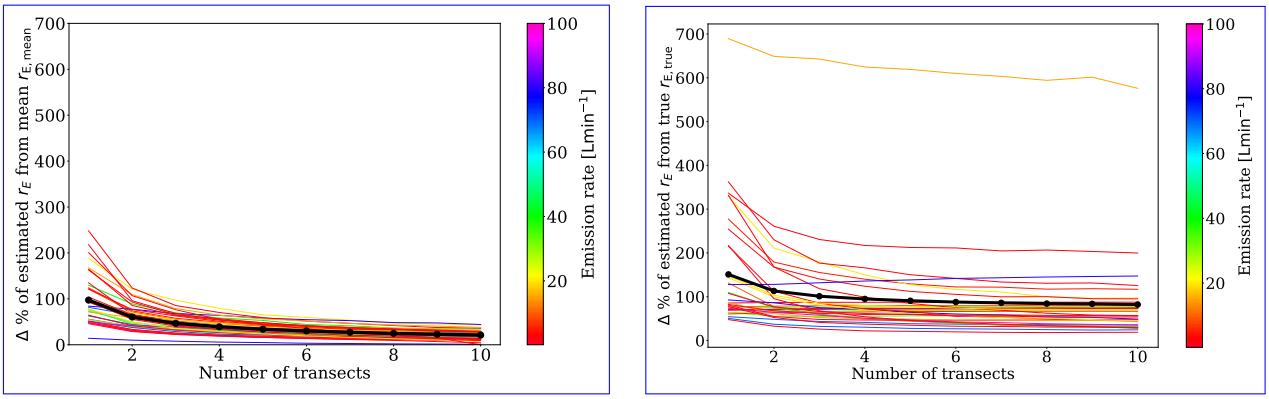
The discussed high uncertainty of individual estimations of the emission rate suggests that the performance may be improved by carrying out several transects. Including more measurements will reduce the random error associated with measurements and increase the confidence to capture the mean of this particular distribution (?)(?). This effect has been demonstrated in ? ? and is illustrated in Fig. 5a. The percentage deviation from the mean at a certain release location and release rateThere, the Monte Carlo mean of the absolute percentage deviation of emission estimation, based on N transects ($N \in [2, 10]$), from the calculated mean emission rate, based on all M transects measured for this release rate, is shown. The percentage deviation from the mean decreases drastically with an increasing number of transects., and converges to 0 for $N = M$ (not shown). Note that here the absolute percentage error is displayed, no difference is made between a negative or positive deviation.

The observation that many of the mean values for individual release rates are still considerably different from the value expected from our conversion equation (Fig. 3) suggests that in addition to random error, there is also a systematic error. This is likely due to two factors; the offset of the sample mean from the population mean (determined by longer time scale weather phenomena) and the offset of the population mean from our linear regression (determined by (built) environment).

Therefore, it follows that the error with respect to the true release rate will not necessarily decrease with more measurements for mean values that are far apart from the regression line. This is the case for the data from London II experiment, which exhibit a large negative offset as well as for the Utrecht I 15 Lmin^{-1} release, which exhibits a very high positive offset (see SI, Sect. S9 S10 for a more detailed discussion). Still, if the offset is not too large, the error in emission estimation decreases with more transects(. This is visualized in Fig. 5b). Note that here the absolute percentage error is displayed, no difference is made between a negative or positive deviation , where the Monte Carlo mean of the absolute percentage deviation of emission estimation in respect to the true emission rate is shown.

The largest decrease in emission estimation error can be achieved from 1 to 3 transects, reducing the estimation error by one third. A further decrease can be reached until around 5-7 transects are included, after which the gained variation reduction levels off, including more transects then reduces the uncertainty by less than 3%.

?



(a) Calculated mean emission rate

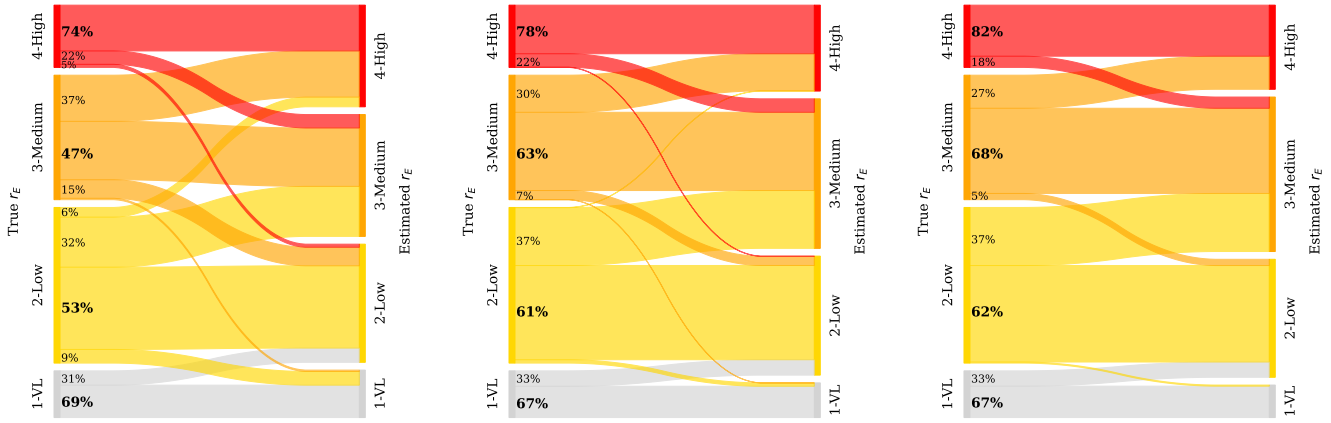
(b) True release rate

Figure 5. Mean absolute percentage deviation of emission estimation from the (a) calculated mean emission rate and (b) true release rate as a function of the number of transects. Each line represents one release rate at one location, and was created by a Monte-Carlo analysis with 2000 repetitions. The total mean over all release rates considered in this analysis is shown in black.

? carried out a similar analysis on multiple passages of real-world leak locations, comparing individual emission rate estimate to the emission rate estimate based on the distribution mean. They reported the sharpest decline in variation with increased detections from 2 to 4, aligning with our analysis. They also noted that, beyond 5-6 transects, the reduction in variability diminishes. Additionally, ~~?~~ **?** reported errors relative to the mean of 50% deviation based on 2 transects. This compares well with the variability in our corresponding analysis, which amounts to an average of 62% for ~~N=2~~ $N = 2$ comparing to the mean (Fig. 5a). Nevertheless, in comparison to true release rates from real measurements, the overall deviations are higher than in the aforementioned studies. We find the mean percentage deviation from the true release rate to be 115% for ~~N=2.~~ $N = 2$ (Fig. 5b). Further, the deviations are skewed, with higher overestimations than underestimations, a result of the right-skewed distribution of measured spatial peak area values.

~~?~~ **?** stated that at least 10 transects are necessary to estimate the source strength within 40% of the true emission, simulating a mobile measurement set-up in an Large Eddy Simulation model. We found for half of the cases an estimation error within 58% when including 10 transects, for the other half errors ranged between 65-200% with one outlier at over 500%. Thus, our findings are in closer agreement with ~~?~~ **?** than with the smaller estimation errors reported in ~~?~~**?**. The generally higher estimation errors demonstrate the complexity of real world environmental influences.

The persistent estimation error, even with the inclusion of more transects, likely reflects the influence of the built-environment and meteorological conditions. If there is a systematic bias for that specific location or time frame of measurements, combining several transects will still include that high or low bias. Additional measurements can only balance out random errors or fluctuations. Taking measurements under different meteorological conditions could potentially improve estimations, enabling sampling of the full distribution of possible realizations of the emission plume.



(a) ~~N=1~~$N = 1$ (b) ~~N=3~~$N = 3$ (c) ~~N=6~~$N = 6$

Figure 6. Categorization performance of the area equation with (a) ~~N=1~~$N = 1$, (b) ~~N=3~~ $N = 3$ and (c) ~~N=6~~ $N = 6$ transects for distributions with a small to medium offset (25 of the 35 release rates). The left axis represents the true r_E, where the width of the bars indicate the amount of plumes belonging to each emission category (categories: 1-Very Low, 2-Low, 3-Medium and 4-High). The right axis represent the categories estimated by the statistical model and the connecting lines visualize the amount of plumes from each category pool which the algorithm classifies into another (or the same) category. The underlying dataset is based on the Monte-Carlo simulation resulting in 2000 estimates per release rate.

Fig. 6 illustrates that the accuracy of categorization on average improves when taking into account one (Fig. 6a), three (Fig. 6b) or six (Fig. 6c) transects. The London II and Utrecht I 15 Lmin^{-1} releases are not included in this analysis because they include strong systematic biases.

For emission rate estimations based on ~~a single detection~~ one single transect ($N = 1$), the categorization performance ranges between 47% and 74%, averaging 58%. When three transects are used ($N = 3$), the categorization performance improves to an average of 65%. Including six transects ($N = 6$) further enhances performance to an average of 69%, ranging from 62% to 82%. More importantly, the number of strong over or underestimations decreases. For example, whereas for only one transect 5% of category 4 peaks and 15% of category 3 peaks are categorized as category 2, this reduces to 0% and 5% respectively for 6 transects. Also the false overestimation of 6% of category 2 peaks into category 4 vanishes when increasing sampling effort to 6 ~~measurements~~transects.

3.5 Suggested Method for Routine Implementation of Leak Surveys

We suggest the following approach for routine mobile leak detection surveys in urban areas:

1. Survey: Perform surveys with an instrument that falls under the '1-1-10 principle' - at least 1 Hz measurement frequency, 1 s residence time and 10 ppb precision. Record GPS data together with the mole fraction measurements, ideally with the same measurement frequency as the CH_4 analyzer. Ensure that the GPS and CH_4 recordings are time-synchronized.

Measure and record the inlet delay time for the CH_4 measurements at the beginning of the survey. Optimally, gather wind measurements as well, such as with an anemometer mounted on the vehicle's roof. If possible, maintain a sufficient distance from previously passed cars while conducting the measurements. During the survey, if an enhancement above the CH_4 background exceeds 0.2 ppm, attempt to conduct six transects downwind of the leak indication.

2. Evaluation: After the survey campaign, determine CH_4 background levels and subtract ~~it~~ them from the CH_4 measurements (CH_4 background at a certain time = 10^{th} percentile of CH_4 mole fractions within a ± 2.5 min time window). Use a peak detection algorithm to identify peaks in the time series, including start and end time of each peak. If available, use additional information such as ethane measurements, methane : carbon dioxide ratio or CH_4 isotopic composition to differentiate between biogenic, combustion and fossil sources. Use GPS data to infer the driving speed to convert the time series to space coordinates and calculate spatial peak areas (using the CH_4 elevation above background levels). In case several transects were done for the same leak indication, take the average of the $\ln([\mathrm{CH_4}]_{\mathrm{area}})$ of the individual peaks detected during the different transects. Calculate the emission rate by inserting the $\ln([\mathrm{CH_4}]_{\mathrm{area}})$ value into the inverse Area eq (Eq. 12).

3. Categorization: Rank the leak into one of the four emission categories (< 0.5 $\mathrm{Lmin^{-1}}$-Very low, $0.5 - 6$ $\mathrm{Lmin^{-1}}$-Low, $6 - 40$ $\mathrm{Lmin^{-1}}$-Medium and > 40 $\mathrm{Lmin^{-1}}$-High) and use this categorization in addition to safety concerns to prioritize leak repair efforts.

4. Report: Report the number of leaks per category and number of leaks per pipeline kilometre. Report the leak size distribution and total emissions for the surveyed area. Ideally, provide these numbers separately for different pipeline material (e.g. cast iron, steel, polyethylene and P.V.C.).

4 Conclusions

Our results clearly highlight systematic differences in maximum CH_4 enhancement recorded by different commercial CH_4 gas analyzers. Maximum enhancements differ by up to a factor ten between instruments. When the spatial peak areas are evaluated, differences between instruments reduce to less than a factor two, confirming our hypothesis that the spatial peak area is a more suitable quantity than the recorded peak maximum. We suggest moving beyond the peak maximum as emission estimation metric to avoid biased emission predictions.

We propose an empirical equation for the emission rate (r_E, in $\mathrm{Lmin^{-1}}$) based on integrated spatial peak area of CH_4 enhancements ($[\mathrm{CH_4}]_{\mathrm{area}}$, in $\mathrm{ppm} * \mathrm{m}$, background levels subtracted from CH_4 measurements):

$$r_E = \exp\left(1.292 \cdot \overline{\ln([\mathrm{CH_4}]_{\mathrm{area}})} - 2.377\right) \tag{13}$$

Here, $\overline{\ln([\mathrm{CH_4}]_{\mathrm{area}})}$ is the mean taken over several transects associated with the same leak indication. This formula is very simple in deployment, it does not require any additional information than the actual CH_4 mole fraction and GPS measurements. ~~But~~However, it therefore ignores other important local influences. The most important parameters are likely the built

environment, wind speed and direction, atmospheric stability and turbulent processes. It is expected that incorporating parameters reflecting those factors could improve emission rate predictions. However, especially when it comes to incorporating turbulent processes, the associated measurement effort would become extensive. Therefore, the strength of our simple statistical model is the simplicity of its implementation, still providing reasonable emission estimates. Higher sampling effort can improve quantification accuracy significantly. Including three transects instead of one improved correct categorization into four emission categories from 47%-74% to 61%-78% and to 62%-82% when incorporating six transects instead of one transect. Thus, applying this method to identify, quantify and prioritize leak repairs in urban natural gas distribution systems can support greenhouse gas emission mitigation during the transition to a fossil-fuel-free energy system.

5 Overview Settings of the Controlled Release Experiments

Controlled release experiments: Overview of release rates, duration of releases per location, number of valid transects and number of valid peaks. In cases where several instruments were mounted on the same vehicle, the number of peaks is higher than the number of transects. The time is given as local time (difference to UTC: London UTC+01:00, London II UTC+01:00, Rotterdam UTC+02:00, Toronto UTC+04:00, Utrecht UTC+01:00, Utrecht II UTC+02:00,). The number of transects/peaks with bike and car platforms are given separately for Toronto Day1 (bike + car).

City Location Release Rate $\left[\mathrm{Lmin}^{-1}\right]$ Duration (local time) Valid Transects Valid Peaks

City Location Release Rate $\left[\mathrm{Lmin}^{-1}\right]$ Duration (local time) Valid Transects Valid Peaks

17011:01 - 15:42 36 7213516:01 - 17:20 13 2613512:16 - 13:05 15 30 17013:22 - 14:11 10 2017016:28 - 17:20 22 4417010:25 - 11:53 42 42170.514:18 - 14:44 31 31150.515:03 - 15:31 29 29130.615:38 - 15:58 29 29110.616:14 - 16:40 24 2415.616:44 - 17:06 26 261117:14 - 17:36 22 22130.617:47 - 18:15 49 491110:08 - 10:47 40 4010.511:02 - 11:32 34 3410.211:37 - 12:23 1 1159:05 - 10:15 49 13811010:15 - 10:58 35 9712010:58 - 11:23 21 6114011:23 - 11:54 34 9718011:54 - 12:44 44 12412013:05 - 13:34 6 24112013:34 - 13:48 3 1214013:48 - 14:26 6 242110:28 - 11:01 21 6220.1511:01 - 11:39 20 5920.51511:39 - 12:12 23 6620.3112:12 - 13:16 22 6633.3313:16 - 14:26 9 3619.916:11 - 16:18 4+4 4+41516:19 - 16:27 5+4 5+412.516:30 - 16:40 5+7 5+7119.816:40 - 16:49 5+5 5+529.99:48 - 9:58 7 72510:03 - 10:11 7 72110:16 - 10:24 11 1120.1210:28 - 10:37 4 420.510:41 - 10:59 19 191313:06 - 13:46 24 4812.1814:22 - 15:17 28 5621513:06 - 13:46 61 12221514:22 - 15:171410:48 - 11:22 10 101411:38 - 11:48 5 511011:48 - 12:18 6 618012:18 - 12:32 6 1112012:32 - 13:01 23 46110014:05 - 14:21 10 2011514:28 - 15:23 32 641416:19 - 17:07 18 3510.1517:07 - 17:44 2 21117:44 - 17:56 1 122.510:50 - 11:31 16 162411:31 - 12:09 12 1220.512:09 - 12:37 13 2020.1512:37 - 13:01 10 2020.314:02 - 14:40 5 1022.214:40 - 15:13 18 362115:13 - 15:52 19 382417:00 - 17:30 17 2226017:30 - 17:38 3 322017:38 - 18:06 11 1128018:06 - 18:25 16 32

5 Instrument Performance: Peak Maximum and Spatial Peak Area

Peak Maximum Spatial Peak Area Comparison of peak maximum (a) and spatial peak area (b) from different instruments in London. Regression fits with intercept 0 are applied to the data for each instrument. The results from the uMEA and LI-7810 analyzers are plotted on the y-axis and the results from the G2301-m instrument on the x-axis. The black dotted line represents the 1:1 line. (Peaks exceeding a maximum of 20 ppm are marked with an 'x' and were excluded from the fitting process.)

Code and data availability. The python code and a sub-sample of the data used to produce the results in this article are available on GitHub: https://github.com/judith-tettenborn/CRE_CH4Quantification.git

Author contributions. Contributed to conception and design: T.R., D.Z.-A., H.M.

Contributed to acquisition of data: J.T., D.S, H.M, C. vd V., A.H., I.V., P. vd B., F.V., L.G., S.A., J.F., D.L., R.F., T.R.

Contributed to analysis and interpretation of data: J.T., D.Z.-A., D.S., H.M., A.H., J.F., T.R.

Drafted and/or revised the article: J.T., D.Z.-A., H.M, A.H., I.V., F.V., L.G., S.A., J.F., D.L., R.F., T.R.

Approved the submitted version for publication: J.T., D.Z.-A., D.S, H.M, C. vd V., A.H., I.V., P. vd B., F.V., L.G., S.A., J.F., D.L., R.F., T.R.

Competing interests. At least one of the (co-)authors is a member of the editorial board of Atmospheric Measurement Techniques.

Disclaimer. Judith Tettenborn was supported through a grant from Environmental Defense Fund.

Acknowledgements. We thank all who contributed to data acquisition during the measurement campaigns across various research groups. Special thanks to Roberto Paglini from Politecnico di Torino and Ceres Woolley-Maisch from Utrecht University for their dedicated efforts during the campaign, assistance with data analysis, and valuable discussions.

In the drafting and programming of this publication the AI tool ChatGPT (https://chat.openai.com/) has been utilized as aid based upon initial drafts. Every response stemming from the AI has been checked, evaluated and only implemented with care.

Supplement Information of

Improving Consistency in Methane Emission Quantification from the Natural Gas Distribution System across Measurement Devices

Judith Tettenborn[1], Daniel Zavala-Araiza[1,2], Daan Stroeken[1], Hossein Maazallahi[1*], Carina van der Veen[1], Arjan Hensen[3], Ilona Velzeboer[3], Pim van den Bulk[3], Felix Vogel[4], Lawson Gillespie[4,5], Sebastien Ars[4], James France[6,7], David Lowry[6], Rebecca Fisher[6], and Thomas Röckmann[1]

[1]Institute for Marine and Atmospheric Research Utrecht (IMAU), Utrecht University, Utrecht, The Netherlands
[2]Environmental Defense Fund, Amsterdam, The Netherlands
[3]Netherlands Organisation for Applied Scientific Research (TNO), Utrecht, The Netherlands
[4]Climate Chemistry Measurements and Research, Climate Research Division, Environment and Climate Change Canada, Toronto, Canada
[5]Department of Physics, University of Toronto, Toronto, Canada
[6]Department of Earth Sciences, Centre of Climate, Ocean and Atmosphere, Royal Holloway, University of London, Egham, United Kingdom
[7]Environmental Defense Fund, London, United Kingdom
[*]Now at: Department of Renewable Energies and Environment, College of Interdisciplinary Science and Technologies, University of Tehran, Tehran, Iran

Correspondence: Thomas Röckmann (t.roeckmann@uu.nl)

Contents

S1 ~~Calculating Residence Time in~~ Instrument ~~Cell~~Characteristics

The ~~residence time~~ G2301 instrument provides atmospheric mole fraction measurements of CH_4 with a data frequency of ≈ 0.36 Hz (every 2.8 s) with a precision of < 0.5 ppb within the operating range of 0-20 ppm.

The G2301-m greenhouse gas analyzer deploys cavity ring-down spectroscopy. It is a modification of the G2301 model designed to minimize effects induced by mobile measurements. It has an acquisition rate of 1 Hz and a precision of <1.5 ppb for CH_4.

The G2401 analyzer has a precision of < 1 ppb for CH_4 over a 5 s integration period.

The G4302 instrument has two operating modes. The one used was the 'ethane/methane' mode, which is characterized by a measurement frequency of > 1 Hz, a precision of 30 ppb in the operating range of 1-5000 ppm. Both instruments utilize cavity ring-down spectroscopy (CRDS) to measure CH_4.

The LI-7810 $CH_4/CO_2/H_2O$ Trace Gas Analyzer is a laser-based gas analyzer that uses Optical Feedback — Cavity-Enhanced Absorption Spectroscopy (OF-CEAS) to detect gases in air. It can measure CH_4 within the range 0-100 ppm with a precision ($1\ \sigma$) of 0.6 (0.25) ppb at 2 ppm with 1 (5) s averaging. Its response time ($T_{10} - T_{90}$ from 0 to 2 ppm is ≤ 2 s.

The MGA10 analyzer measured at 1 Hz with precision of 1 ppb within the measurement range 0-200 ppm.

The Mira Ultra instrument has a measurement frequency of 1 Hz, a sensitivity of < 2 ppbs^{-1} and an operation range 0.02-10,000 ppm. The temporal response is 1 s and it takes 3 s to 90 % recovery with it's internal pump. It deploys a mid-infrared laser absorption spectroscopy technology.

The TILDAS Dual Laser Trace Gas Analyzer measured at 1 Hz with precision of 2.4 ppb and had a response time equal to about 2 s.

The UGGA device has a precision of < 2 ppb for CH_4 over a 1 s integration period and its measurement range lies between 0.01-100 ppm.

The uMEA analyzer uses laser absorption spectroscopy, delivering linear measurements within the range 0.01-100 ppm and has a precision of 3 ppb for CH_4 over a one second period.

The control range of the Alicat mass flow controller is 0-100 Lmin^{-1} under standard conditions with a measurement accuracy

of ±(0.8% of reading + 0.2% of full scale).

The residence time within the measurement cell for the different instruments was determined on the basis of the cell temperature T_{cell} [K], cell pressure p_{cell} [Pa], cell volume V_{cell} [m^3] and flow rate Q_{cell} [slm] specified by the manufacturers (given in Tab. S1). The units in brackets specify the units in which the different quantities have to be inserted into the ~~equation~~equations Eq. S1 and Eq. S2.

The normalized volume (scaled to standard pressure 101325 Pa and standard temperature 25°C) was calculated:

$$V_{\mathrm{norm}} = \frac{p_{\mathrm{cell}} \cdot V_{\mathrm{cell}} \cdot R \cdot T_{\mathrm{norm}}}{R \cdot T_{\mathrm{cell}} \cdot p_{\mathrm{norm}}} \tag{S1}$$

Then, given the flow rate, the residence time was determined from V_{norm} and the flow rate Q_{cell} as:

$$\tau = \frac{V_{\mathrm{norm}}}{Q_{\mathrm{cell}}} \tag{S2}$$

Table S1. Overview of instrument characteristics of analyzers deployed in the controlled release experiments. In cases where flow rate varied, the bold numbers were used for the calculation of the residence time. The integration time over which the precision applies was not available for some instruments.

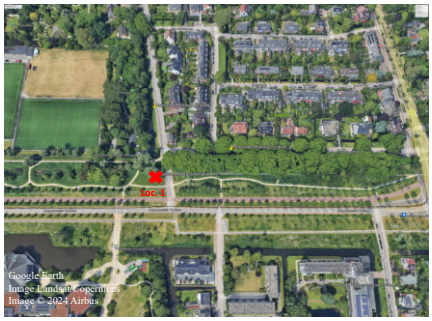
heightGHG Analyzer	T_{cell} [°C]	p_{cell} [mbar]
heightG2301[a]	45	190
G2401[a]	35	186
~~UGGA[b]~~ G4302[a]	~~25~~ 35	~~186~~ 600
LI-7810[c] ~~[b]~~	55	390
~~TILDAS[d]~~ MGA10[c]	~~25~~ 27	~~40~~ 80
Mira Ultra[e] ~~[d]~~	42	240
~~MGA10[f]~~ TILDAS[e]	~~27~~ 25	~~80~~ 40
UGGA[f]	25	186
uMEA[f]	.	.

[a]Picarro INC, Santa Clara, USA. [b]LI-COR Environmental, Lincoln, USA. [c]MIRO Analytical AG, Wallisellen, CH. [d]Aeris Technologies, Eden Landing Road Hayward, CA. [e]Aerodyne Research, Billerica, USA. [f]Los Gatos Research, San Jose, USA.

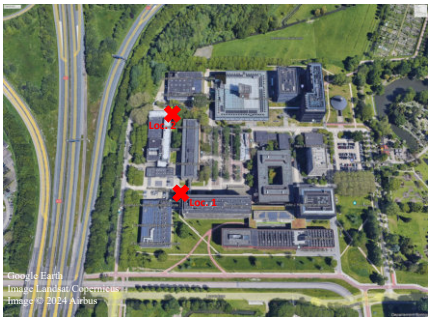
height

S2 Description Controlled Release Experiments

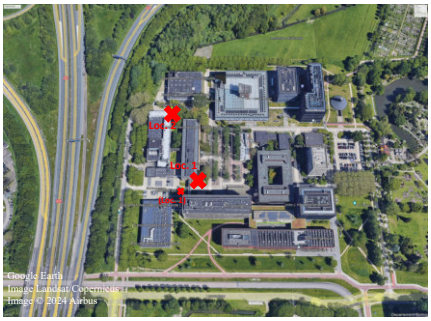
S2.1 Release Locations



(a) Rotterdam



(b) Utrecht I

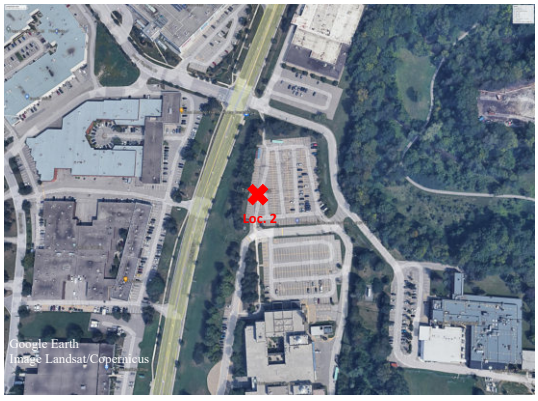


(c) Utrecht II

(d) London I

(e) London II

(f) Toronto Loc. 1



(g) Toronto Loc. 2

Figure S1. Google Earth screenshots of locations of the different controlled release experiments (Google Earth, Image Landsat/Copernicus and Image ©2024 Airbus and Image NOAA). The red crosses indicate the location of the controlled CH_4 releases.

(a) UUAQ car (b) Location 1 - release (c) Location 1 - gas vessel

Figure S2. Rotterdam: Overview of measurement set-up.

(a) Location 1 (b) Location 2

Figure S3. Utrecht II: Overview of measurement set-up.

S2.2 ~~Procedure of the Controlled~~ Release Rates, Detection Counts, and Timing Overview

~~Rotterdam~~

The control range of the Alicat mass flow controller is 0-100 under standard conditions with a measurement accuracy of ±(0.8% of read

City

London I Day1 (September 10, 2019)

London I Day2 (September 11, 2019)

London I Day3 (September 13, 2019)

London II Day1 (May 13, 2024)

London II Day2 (May 14, 2024)

The Mira Ultra instrument has a measurement frequency of

~~The G4302 instrument has two operating modes. The one used was the 'ethane/methane' mode, which is characterized by a measuremer~~

City

Rotterdam (September 6, 2022)

~~**Utrecht**~~

~~Utrecht I~~

~~CH_4 was released simultaneously from two cylinders at two different locations. Two manual flowmeters (Krohne DK800/PV (25-250 N~~

Toronto Day1 (October 20, 2021)

Toronto Day2 (October 24, 2021)

Utrecht I (November 25, 2022)

~~The Mira Ultra instrument has a measurement frequency of~~

~~The G4302 instrument has two operating modes. The one used was the 'ethane/methane' mode, which is characterized by a measuremen~~

City

~~Utrecht II~~

~~Initially, the same two release locations from the previous experiment were used. However, after encountering power supply issues with~~

Utrecht II (June 11, 2024)

~~**London**~~

~~London I~~

~~The LI-7810 CH_4/CO_2/H_2O Trace Gas Analyzer is a laser-based gas analyzer that uses Optical Feedback — Cavity-Enhanced Absorpti~~

~~London II~~

~~The same LI-7810 CH_4/CO_2/H_2O Trace Gas Analyzer as in the previous campaign was utilized.~~

~~**Toronto**~~

~~On~~

S3 Raw Data Processing

The raw measurements taken by the G4302, G2301 and Mira Ultra instruments during the Rotterdam~~and Utrecht~~, Utrecht I and Utrecht II controlled releases were corrected utilising calibration equations obtained by calibration measurements in the IMAU laboratory. The data collected by the other CH_4 analyzer were treated and calibrated by the team that deployed them.

G2301:

$$[CH_4]_{\text{calibrated}} = 1.03127068196 \cdot [CH_4]_{\text{raw}} - 0.15799666857 \tag{S3}$$

G4302:

$$[CH_4]_{\text{calibrated}} = 1.01924906721 \cdot [CH_4]_{\text{raw}} - 0.05887406866 \tag{S4}$$

Mira Ultra:

$$[CH_4]_{\text{calibrated}} = 1.01354227768 \cdot [CH_4]_{\text{raw}} - 0.05055326961 \tag{S5}$$

$[CH_4]$ refers to the CH_4 mole fraction in ppm.

S4 Overview Time Series

Fig. S4 to Fig. S9 show an overview of selected timeseries. The different release rates translate into different peak heights over time. The methane mole fractions measured by different instruments differ strongly, even though the instruments transect the CH_4 plume simultaneously and draw air from the same inlet.

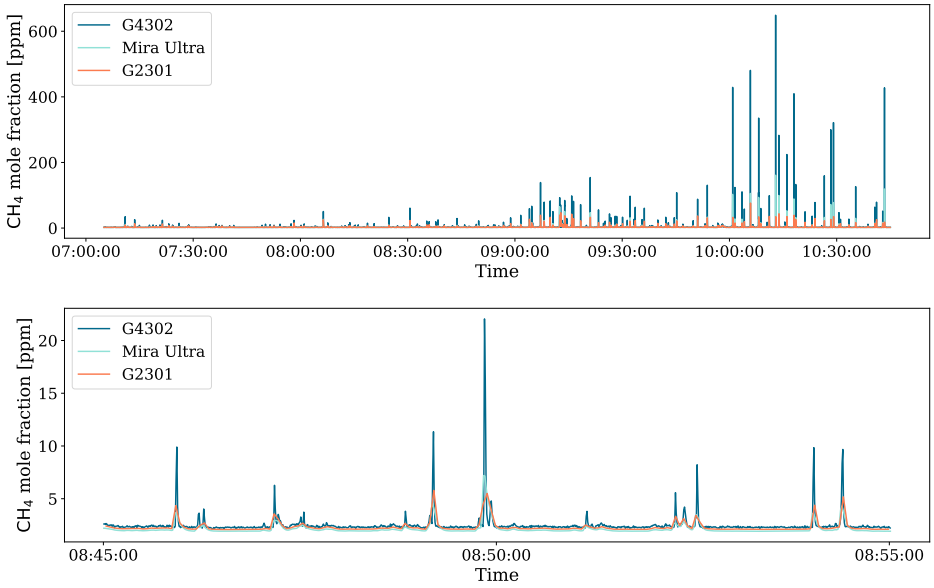


Figure S4. Rotterdam: Timeseries of CH_4 mole fraction, obtained by the G4302, G2301 and Mira Ultra ~~device~~devices, while installed in the UUAQ car. The lower panel displays a zoom to a 10 min measurement interval. Time displayed in UTC.

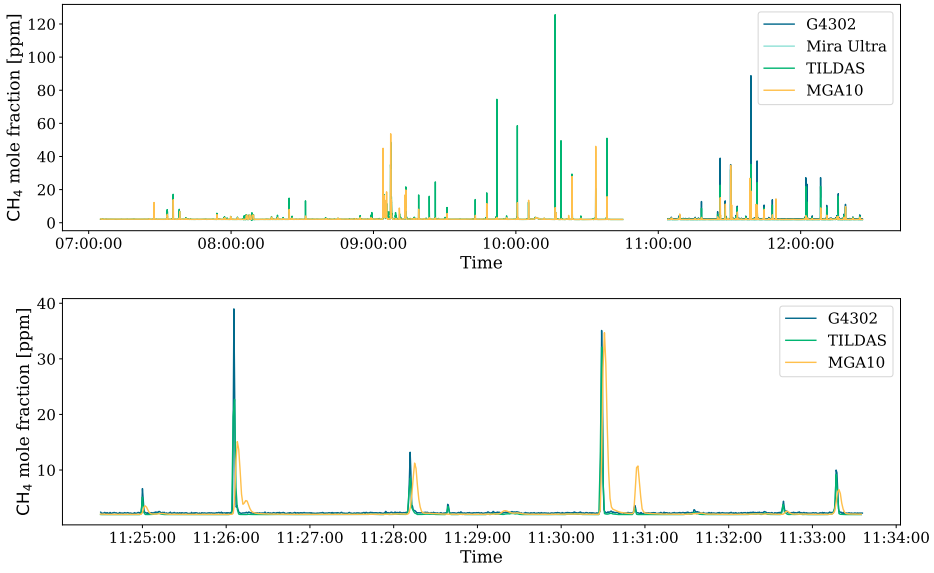


Figure S5. Rotterdam: Timeseries of CH_4 measurements, obtained by the MGA10, TILDAS~~and~~, G4302 and Mira Ultra devices, while installed in the TNO truck. The lower panel displays a zoom to a 10 minutes measurement interval. Time displayed in UTC.

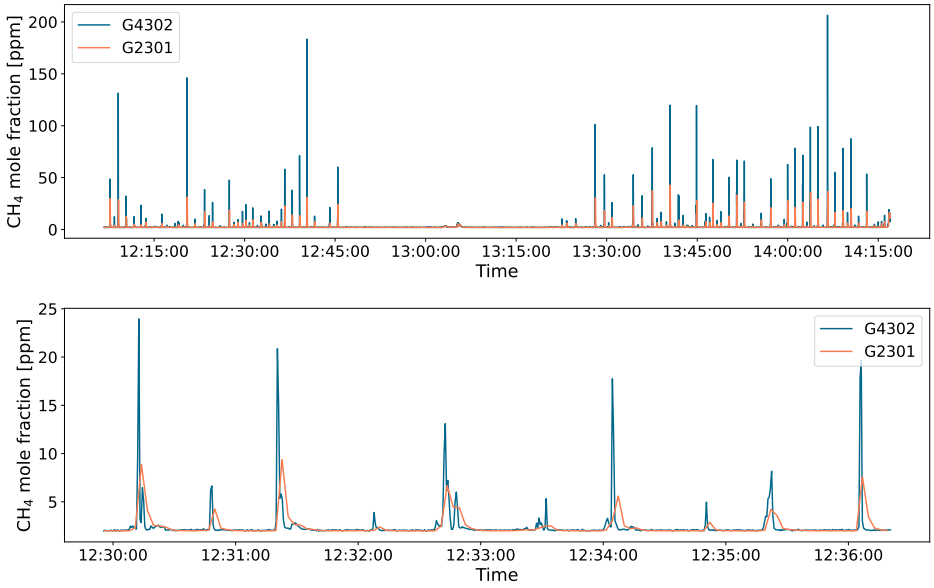


Figure S6. Utrecht I: Timeseries of CH_4 measurements, obtained by the G4302 and G2301 devices. The lower panel displays a zoom to a 6 minutes measurement interval. Time displayed in UTC.

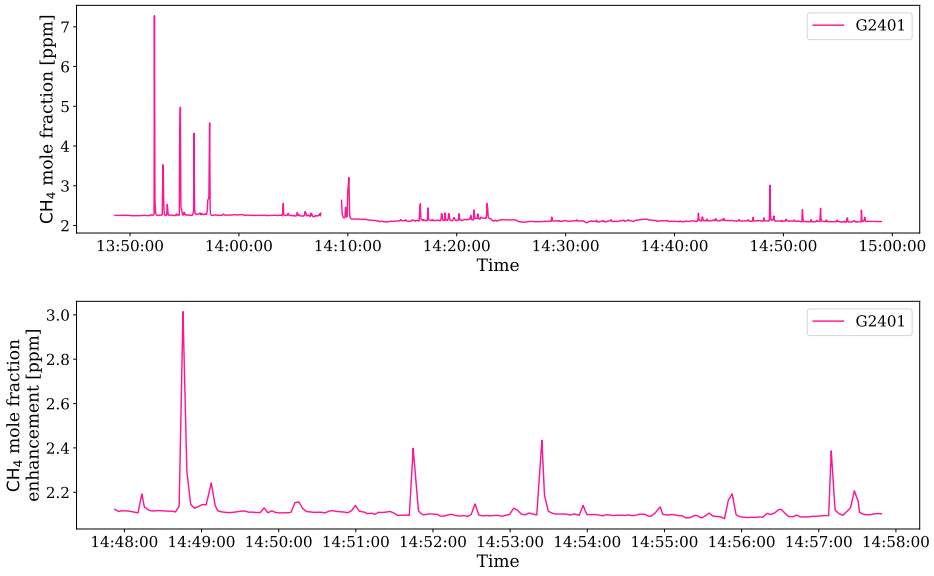


Figure S7. Toronto Day 2 - car: Timeseries of CH_4 measurements, obtained by the G2401 device. The lower panel displays a zoom to a 10 minutes measurement interval. Time displayed in UTC.

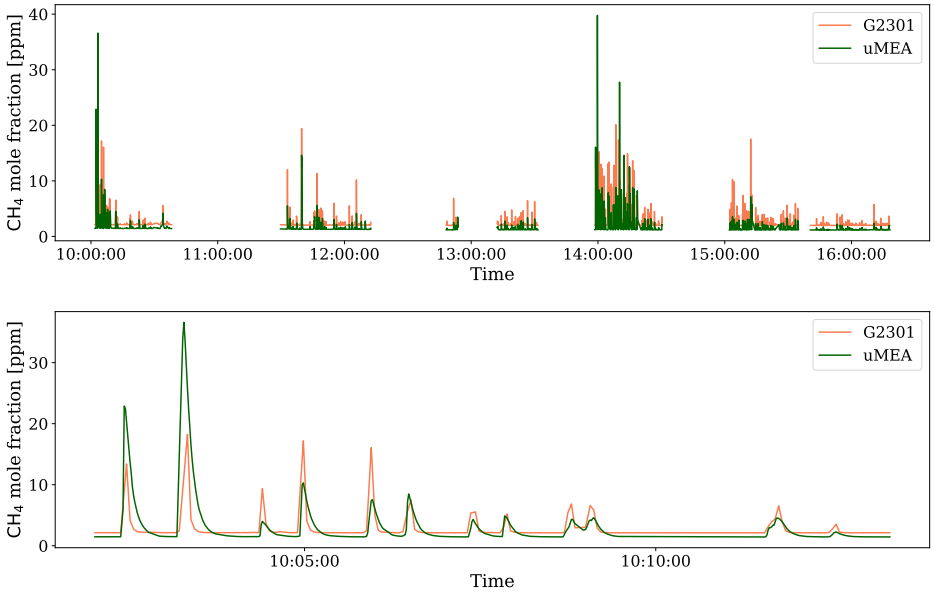


Figure S8. London I Day 1: Timeseries of CH_4 measurements, obtained by the G2301 and uMEA devices. The lower panel displays a zoom to a 10 minutes measurement interval. Time displayed in UTC.

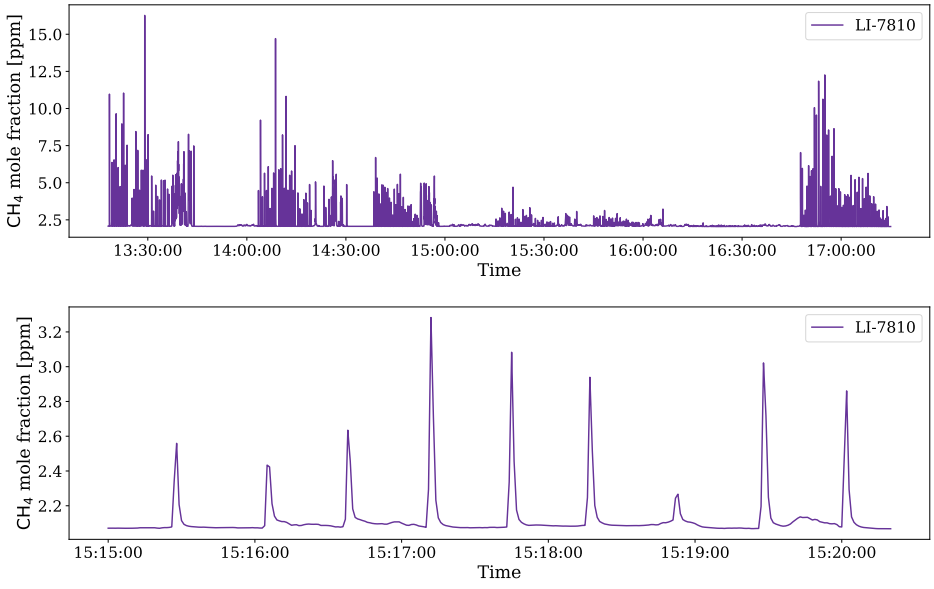
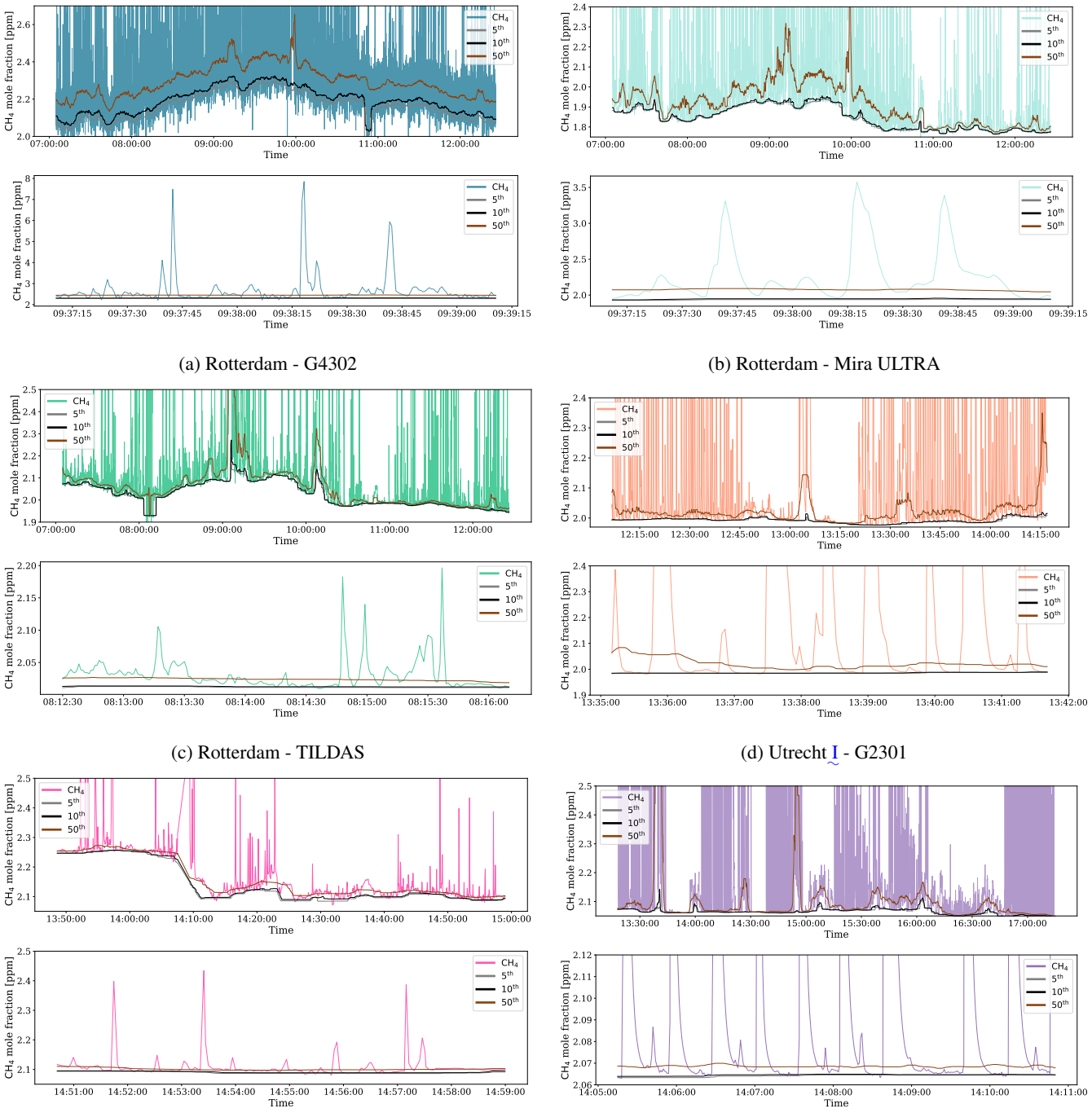


Figure S9. London II Day 1: Timeseries of CH_4 measurements, obtained by the LI-7810 device. The lower panel displays a zoom to a 5 minutes measurement interval. Time displayed in UTC.

S5 Background ~~comparison~~Comparison

Different background mole fraction definitions are used in the literature, using either a fixed threshold or a dynamic one, which offer the advantage to take temporal or spatial variability in the background level into account ~~(von Fischer et al. (2017))~~(von Fischer et al., 2017). Commonly, a moving window is applied and the background is defined as a specific percentile of the data range. Different percentiles were used in the previous literature to set the background, ranging from the 5^{th} percentile in ~~Ars et al. (2020)~~ Ars et al. (2020) to the 50^{th} percentile (median) in ~~Weller et al. (2018)~~ Weller et al. (2018) or taking the mean in ~~von Fischer et al. (2017)~~von Fischer et al. (2017). Higher percentiles will be more strongly influenced by high CH_4 mole fractions when transecting a plume. The mean will be even more distorted towards higher values than the median. This can lead to high background mole fractions which do not represent the ambient background, but are artefacts of a spatially extended CH_4 plume. In this study, the background was defined as the 10^{th} percentile of the CH_4 mole fractions, which was assessed to represent the background well (Fig. S10). The 50^{th} percentile was too strongly influenced by the CH_4 release, occasionally showing up to 0.3 ppm higher background mole fractions compared to the 10^{th} percentile.



(a) Rotterdam - G4302

(b) Rotterdam - Mira ULTRA

(c) Rotterdam - TILDAS

(d) Utrecht I - G2301

(e) Toronto - G2401

(f) London II - LI-7810

Figure S10. Comparison of different background concentrations, determined using three different threshold levels (5^{th}, 10^{th} and 50^{th} percentile). The y-axis is truncated to enhance readability.

S6 Distance Analysis

The left panel in Fig. S11 visualizes the logarithm of the spatial peak area ~~values~~ measured per transects from all controlled release experiments as function of the distance to the CH_4 release location. The right panel of Fig. S11, along with both panels in Fig. S12–Fig. S14 presents the same relationship separately for the different controlled release experiments, with the various release rates distinguished by colour. To evaluate the nature of this relationship, a linear regression was fitted to the log spatial peak area and distance values for each release rate. The spatial peak area values generally decrease with increasing distance, though the effect varies across cases and is relatively minor in most cases within a 75 m range. At some instances, ~~the~~ e.g. for the 5 Lmin^{-1} release in Rotterdam, the linear regression fit even shows a positive slope, suggesting an increase in spatial peak area values with distance. This could be due to the small sample size and the ~~high~~ strong influence of noise~~.~~, such as changing winds or turbulence. For example, if turbulent motions cause the plume to diffuse more strongly (both horizontally and vertically) at a given moment, a transect passing nearby may measure a smaller spatial peak area than a more distant transect, where the wind conditions allowed the plume to remain relatively compact with minimal diffusion. Overall, these findings suggest that distance may not be a major factor affecting peak detection in urban areas, where peaks are expected to be identified primarily within a 75 m range from the source.

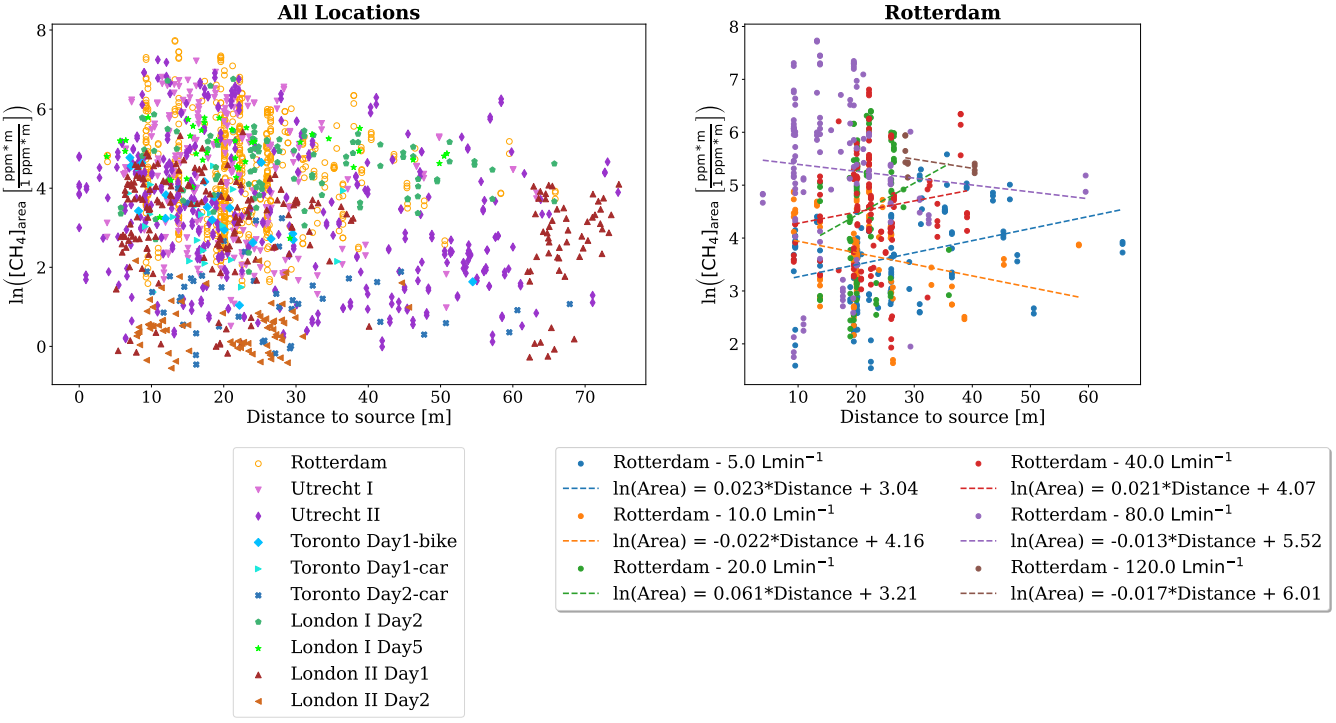


Figure S11. Logarithmic spatial peak area as function of distance to source for all individual CH_4 enhancements reported in this study (a) and Rotterdam (b). In panel (b), colours represent different release rates, with a separate linear regression fitted for each rate.

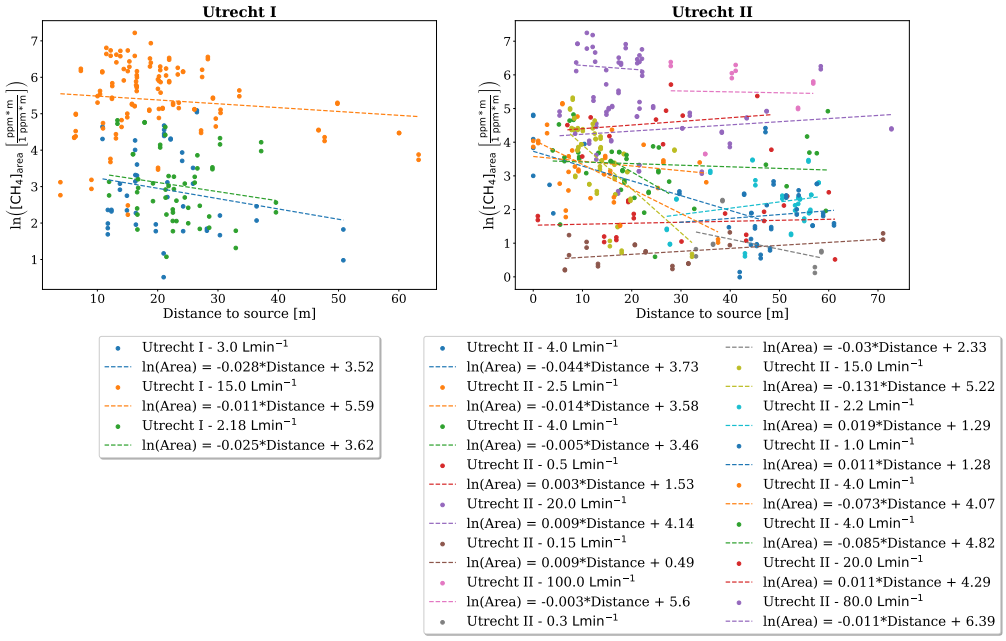


Figure S12. Logarithmic spatial peak area as function of distance to source for individual CH_4 enhancements reported in the Utrecht I (a) and Utrecht II (b) controlled release experiment. Colours represent different release rates, with a separate linear regression fitted for each rate.

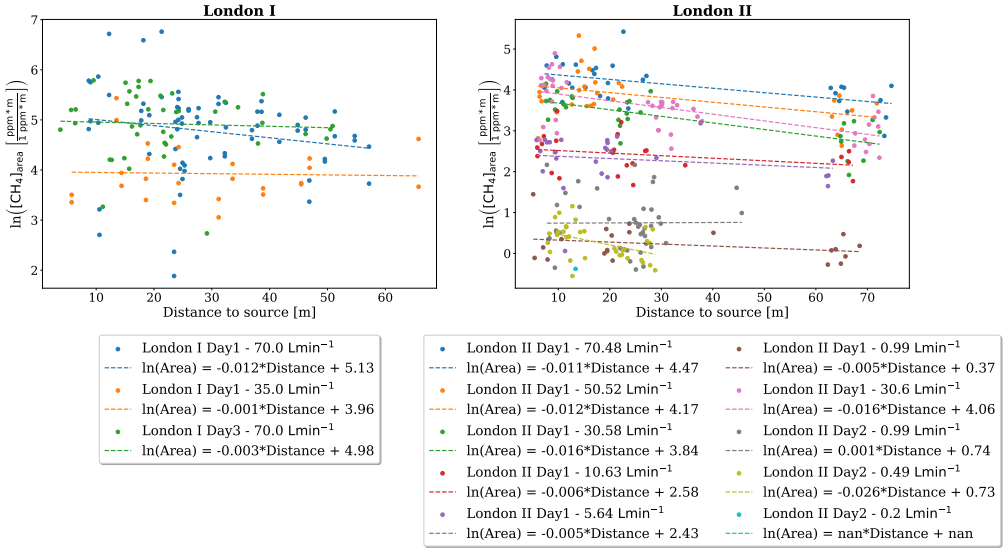


Figure S13. Logarithmic spatial peak area as function of distance to source for individual CH_4 enhancements reported in the London I (a) and London II (b) controlled release experiment. Colours represent different release rates, with a separate linear regression fitted for each rate.

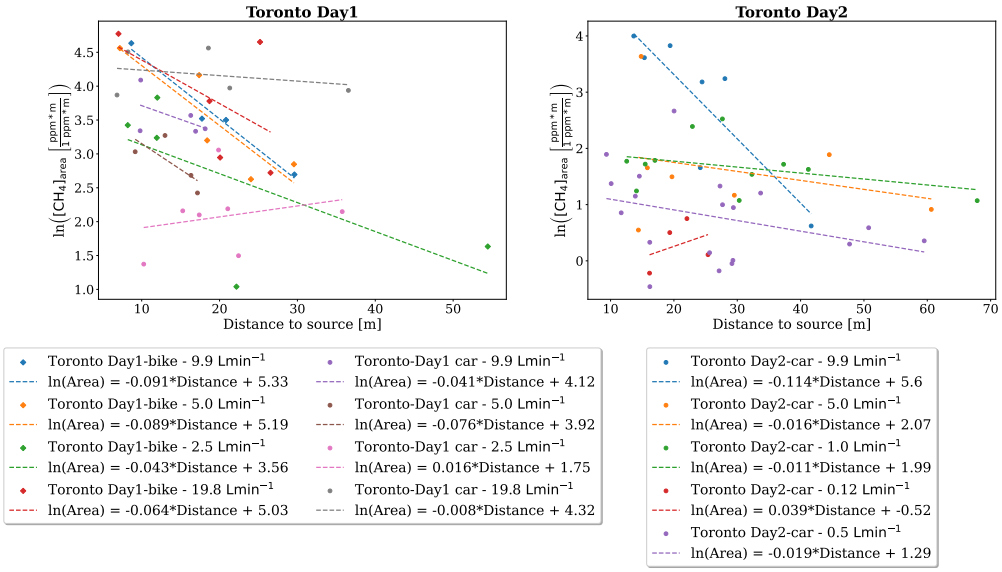


Figure S14. Logarithmic spatial peak area as function of distance to source for individual CH_4 enhancements reported in the ~~Toronto Day 1~~ (a) Toronto Day1 and ~~Day 2~~ (b) Day2 controlled release experiment. Colours represent different release rates, with a separate linear regression fitted for each rate.

S7 Model Diagnostics

There are four main assumptions underlying a linear regression model which describes the relation of a response variable Y and a predictor variable X ~~(Von Storch and Zwiers (2002), Flatt and Jacobs (2019))~~(Von Storch and Zwiers, 2002; Flatt and Jacobs, 2019) :

1. Linearity: The relationship between X and the mean of Y is linear.
2. Homoscedasticity: The variance of residuals is the same for any value of X.
3. Independence: Observations are independent of each other.
4. Normality: For any fixed value of X, the error terms (residuals) of Y are normally distributed.

Violations of these assumptions can lead to biased and misleading inferences, confidence intervals, and scientific insights ~~(Flatt and Jacobs (2019))~~(Flatt and Jacobs, 2019).

S7.1 Analysis of Residuals

To judge on linearity, it can be helpful to visualize the shape of the residuals. This can be done via a standardized residuals plot, where systematic behaviour can be assessed ~~(Von Storch and Zwiers (2002), Biecek and Burzykowski (2021))~~(Von Storch and Zwiers, 200

. Standardized residuals are the differences between the observed values and the values predicted by the fitted linear regression model, divided by the standard deviation of the error estimates:

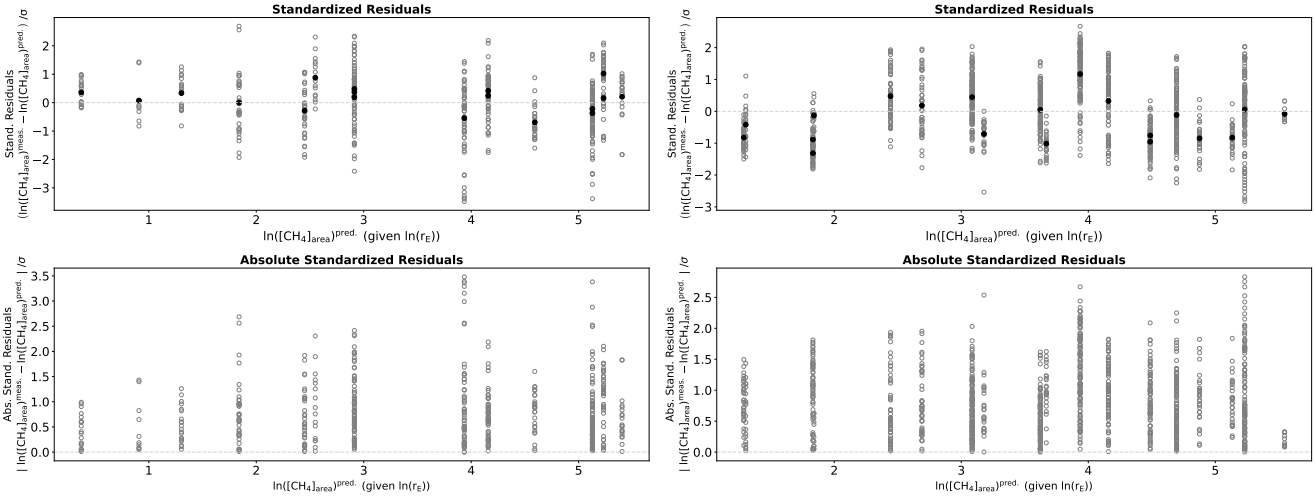
$$\frac{\ln([CH_4]_{area})^{measured} - \ln([CH_4]_{area})^{predicted}}{\sigma} \tag{S6}$$

They are plotted against the estimated conditional mean $\mu_{\mathbf{Y}_i|\mathbf{X}=\mathbf{x}_i}$, i.e. the values predicted by the regression (in this case $\ln([CH_4]_{area})$) for the given values of the independent variable ($\ln(r_E)$). Homoscedasticity means errors e_i all have common variance. Violations of this can influence the coefficients derived under ordinary least-squares regression. Scatter plots of the absolute residuals can help detecting heteroscedasticity. The third assumption necessitates observations to be independent of each other. Paired samples represent the most basic example of non-independent data. When data fail to satisfy the independence assumption, it can impair the accuracy of test statistics ~~(Nimon (2012))~~(Nimon, 2012). In a good fitting model, residuals should exhibit random, not systematic deviations from zero. This entails their distribution being symmetric around zero (mean should be zero). Additionally, residuals should have minimal variability, ideally being close to zero themselves ~~(Biecek and Burzykowski (2021))~~(Biecek and Burzykowski, 2021). Normality can be assessed using quantile-quantile plots (QQ plot) or test statistics, whereby the Shapiro-Wilk test was found to be the most powerful tests in most situations ~~(Keskin (2006), Razali and Wah (2011))~~(Keskin, 2006; Razali and Wah, 2011). Here, the Shapiro-Wilk and the Lilliefors test (a modification of the Kolmogorov-Smirnov test) were applied to the residuals, using a 5% significance level. This was done utilizing the *scipy.stats module* (*stats.shapiro*) and the *statsmodels.stats.diagnostic* module (*lilliefors*).

Fig. S15 illustrates the standardized residuals for the area linear regression model. The x-axis displays the predicted values of $\ln([CH_4]_{area})^{predicted}$), i.e. the vertical point clouds represent the different release rates, but plotted here in terms of the corresponding $\ln([CH_4]_{area})$ estimate based on the Area eq. For visibility, the different releases were plotted in two groups and only distribution with at least 10 observations are shown.

The majority of the means (indicated as a black dot) fall relatively close to the zero line. There is a small tendency towards negative deviations from zero for the means. The residuals do not scatter symmetrically around their mean for all distributions. Clustering of data towards the center can be indicative of a normal distribution. This seems to be the case e.g. for the release rate 40 $Lmin^{-1}$ (Fig. S15b, $\ln([CH_4]_{area}) = 4.7$), where also the mean is close to zero. However, other distributions of residuals are more scattered. Some distributions show long tails, suggesting skewness in the distribution, e.g. residuals at 15 $Lmin^{-1}$ in Utrecht I or 80 $Lmin^{-1}$ in Rotterdam (Fig. S15b, corresponding to an $\ln([CH_4]_{area})$ estimate of 3.9 and 5.2 respectively).

The absolute standardized residuals predominantly remain below 3, and mostly even under 2. There is a weak trend of increasing residual variability with higher release ~~rtaes~~rates. Nonetheless, this trend is marked by significant fluctuations (Fig. S15b, lower panels).



(a) Utrecht II and London I

(b) Rotterdam, Utrecht I, London II, Toronto

Figure S15. Standardized residuals (differences between the measured values and the values predicted by the fitted line, divided by the standard deviation of the error estimates) plotted against the conditional estimate $\ln([\mathrm{CH_4}]_{\mathrm{area}})_i^{\mathrm{pred.}} \mid X = \ln(\mathrm{r_E})_i$ for the linear regression (upper panel). The lower panel displays the absolute values of these standardized residuals. For better visibility, the dataset was separated into (a) Utrecht II and London I and (b) Rotterdam, Utrecht I, London II, Toronto data.

S7.2 Statistical Normality Tests

The results (pass or fail, p-values and statistics) of the Shapiro-Wilk (SW) and the Lilliefors test are provided in Tab. S3 to Tab. S7. Data with small sample size were omitted from this analysis.

From the 6 assessed release rates for Rotterdam, 4 passed the Shapiro-Wilk test and even 3 passed the Lilliefors test, which means the hypothesis that the data follow a normal distribution could not be rejected in those cases (Tab. S3). It was rejected however in both tests for the release rate of 20 $\mathrm{Lmin^{-1}}$ and 80 $\mathrm{Lmin^{-1}}$. Despite having high test statistics for the SW test, the corresponding p-values are low. Only the 2.18 $\mathrm{Lmin^{-1}}$ release in Utrecht I passed both tests (Tab. S4). The p-value for the 3 $\mathrm{Lmin^{-1}}$ release, which passes the SW test, is 0.039 for the Lilliefors test, so comparably close to 0.05, therefore only narrowly failing. In Utrecht II 6 of the 10 distributions pass the SW test and 6 the Lilliefors test (Tab. S5). For the London ~~I~~ CREs, two out of three experiments pass the Lilliefors normality test, while only one out of three passes the Shapiro-Wilk test (Tab. S6). For a release rate of 35 $\mathrm{Lmin^{-1}}$, both tests indicate normality. For the 70 $\mathrm{Lmin^{-1}}$ release rate, the outcomes differ between the two tests and experiment days. In London II 7 of the 9 releases pass the SW test, while all pass the Lilliefors test (Tab. S7).

Overall, in most cases half or the majority of distributions passes the normality tests. This means on the other side that a significant number of distributions do not pass. The statistic values from the Lilliefors test are generally lower compared to the Shapiro-Wilk test, which may suggest that the Lilliefors test is less sensitive to deviations from normality in these specific datasets.

Table S3. Rotterdam: Normality statistics summary.

Release Rate	Dataset Size	Shapiro-Wilk Test			Lilliefors Test		
[Lmin^{-1}]		Result	p-value	Statistic	Result	p-value	Statistic
5	138	pass	0.367	0.989	pass	0.707	0.047
10	97	pass	0.547	0.988	pass	0.441	0.064
20	85	fail	0.0	0.918	fail	0.002	0.133
40	121	pass	0.782	0.993	pass	0.730	0.049
80	124	fail	0.0	0.956	fail	0.017	0.093
120	12	pass	0.057	0.865	fail	0.016	0.273

Table S4. Utrecht I: Normality statistics summary.

Release Rate	Dataset Size	Shapiro-Wilk Test			Lilliefors Test		
[Lmin^{-1}]		Result	p-value	Statistic	Result	p-value	Statistic
2.18	56	pass	0.168	0.97	pass	0.613	0.076
3	48	pass	0.078	0.957	fail	0.039	0.132
15	122	fail	0.0	0.950	pass	0.073	0.08

Table S5. Utrecht II: Normality statistics summary.

Release Rate	Dataset Size	Shapiro-Wilk Test			Lilliefors Test		
[Lmin^{-1}]		Result	p-value	Statistic	Result	p-value	Statistic
0.15	29	fail	0.002	0.865	pass	0.198	0.135
0.5	20	pass	0.412	0.953	pass	0.246	0.153
1	39	pass	0.246	0.964	pass	0.690	0.084
2.2	36	pass	0.385	0.968	pass	0.628	0.091
2.5	16	pass	0.879	0.973	pass	0.85	0.111
4	79	pass	0.116	0.975	pass	0.07	0.1
15	70	fail	0.0	0.926	fail	0.001	0.152
20	67	pass	0.067	0.966	fail	0.039	0.116
80	46	fail	0.003	0.918	fail	0.008	0.155
100	28	fail	0.002	0.863	fail	0.001	0.222

Table S6. London I: Normality statistics summary.

Release Rate	Dataset Size	Shapiro-Wilk Test			Lilliefors Test		
[$\mathrm{Lmin^{-1}}$]		Result	p-value	Statistic	Result	p-value	Statistic
35	60	pass	0.067	0.963	pass	0.117	0.106
70	114	fail	0.0	0.956	fail	0.017	0.096
70	42	fail	0.004	0.913	pass	0.147	0.119

Table S7. London II: Normality statistics summary.

Release Rate	Dataset Size	Shapiro-Wilk Test			Lilliefors Test		
[$\mathrm{Lmin^{-1}}$]		Result	p-value	Statistic	Result	p-value	Statistic
0.49	34	pass	0.879	0.984	pass	0.751	0.087
0.99	40	pass	0.254	0.965	pass	0.461	0.096
0.99	22	fail	0.016	0.886	pass	0.107	0.168
5.64	26	fail	0.0	0.827	pass	0.129	0.152
10.63	24	pass	0.713	0.972	pass	0.679	0.106
30.58	30	pass	0.054	0.932	pass	0.174	0.136
30.6	51	pass	0.622	0.982	pass	0.749	0.071
50.52	29	pass	0.537	0.969	pass	0.463	0.112
70.48	31	pass	0.725	0.977	pass	0.286	0.122

S7.3 Spatial Peak Area distribution per Release Rate

Fig. S16 to Fig. S20 provide an overview of the spatial peak area distributions per release rate in the form of histograms and quantile-quantile (QQ) plots. For each histogram, a Gaussian distribution is plotted together with the data, employing mean and standard deviation derived from the underlying dataset. In the QQ plots, the vertical axis displays the ordered logarithmic spatial peak area values, while the horizontal axis displays expected values based on the standard normal distribution. When the normality assumption is met, the plot should exhibit points scattered closely along the 45-degree diagonal line. While the normality assumptions must be met by the residuals, here the $\ln([\mathrm{CH_4}]_{\mathrm{area}})$ values are plotted for easier comparison with Figure 3 in the main manuscript. Since the residuals for each release rate are obtained by subtracting a scalar from the $\ln([\mathrm{CH_4}]_{\mathrm{area}})$ distribution, the distribution's shape remains unchanged and is simply shifted by this scalar.

For Rotterdam the histograms for the 20, and 80 $\mathrm{Lmin^{-1}}$ releases appear to exhibit a bimodal shape (Fig. S16a). This is also reflected in the QQ plots of the 20 and 80 $\mathrm{Lmin^{-1}}$ release rates (Fig. S16b). Variations are observed in the central body

of the 80 Lmin^{-1} release, and more pronounced deviations are evident in the case of the 20 Lmin^{-1} release, which exhibits an s-shaped pattern. This visualizes why the normality tests fail. For the other releases (except 120 Lmin^{-1}, for which the low number of data points makes an analysis difficult) the distribution aligns well with the 1:1 line in the QQ plots. In all instances, the highest quantiles consistently appear below the $45°$ line, indicating a scarcity of data in the high range compared to a normal distribution (a thinner tail on the right side). For some cases, the points also fall below the 1:1 line for the lowest quantiles, implying a higher abundance of data at the low range compared to a normal distribution (a fatter tail on the left side).

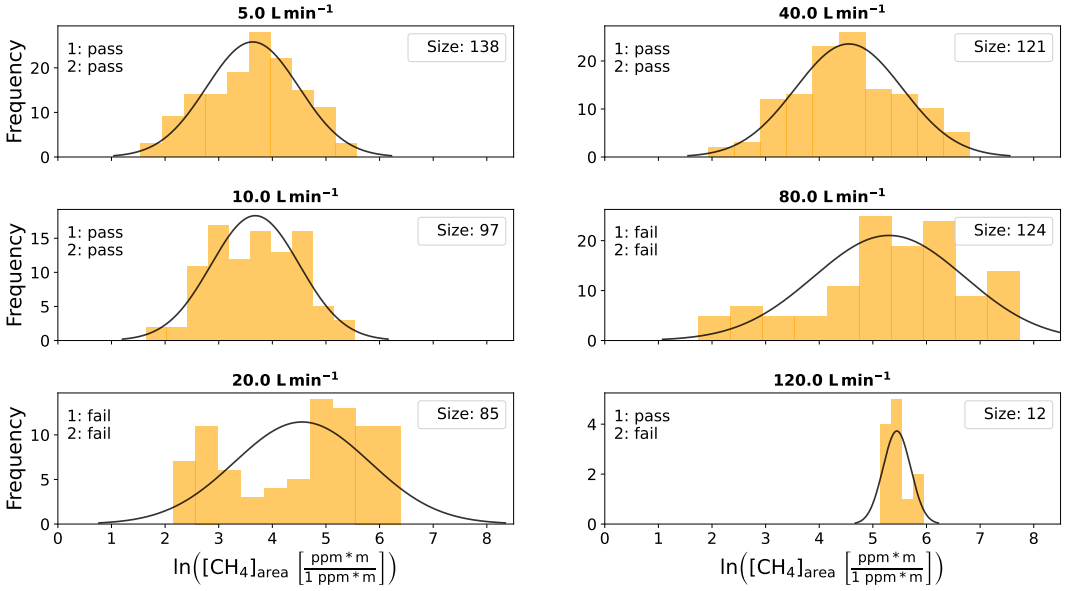
Both the histogram and QQ plot of the 2.18 Lmin^{-1} release in Utrecht I confirm the positive assessments of both normality tests (Fig. S17). However, the 3 Lmin^{-1} QQ plot exhibit an s-shaped form, confirming the fat tails visible in the histogram plot. The 15 Lmin^{-1} release rate distribution shows a skew towards higher $\ln([\mathrm{CH_4}]_{\mathrm{area}})$ values (left-skewed), visible by its concave curve in its QQ plot, explaining the rejection of normality by the SW test.

In the Utrecht II dataset, the right skewed distribution of the 0.15 Lmin^{-1} release could be caused by the peak detection threshold, cutting of part of the distribution. The three releases which fail both tests (15, 80 and 100 Lmin^{-1}) show a bimodal disribution, which appears as s-shape in the QQ plot (Fig. S18).

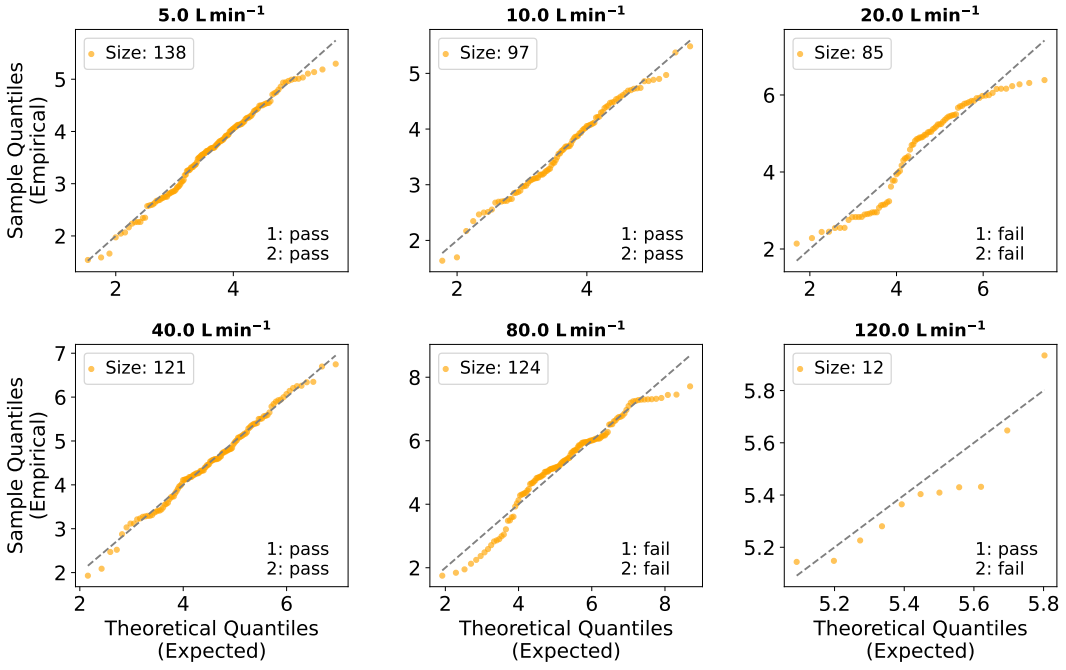
The London I Day2-70 Lmin^{-1} release exhibits a left-skewed distribution according to both the histogram and QQ plot (departing in negative direction from the 1:1 line for both margins) and fails both tests (Fig. S19). The QQ plot for the Day3-70 Lmin^{-1} release suggests normality, similar to the Lilliefors test, only disturbed by two outliers, which could be the reason why the SW test failed.

Similar to the good performance of the distributions of the London II data in the two test statistics, the visual observation of the histogram and QQ plots also shows normality in almost all cases (Fig. S20). The Day2-5.64 Lmin^{-1} release does not show large deviations in the QQ plot but exhibits an outlier which likely causes the SW test to fail.

Generally, as the release rates increase, a shift of the centre of the distributions towards higher $\ln([\mathrm{CH_4}]_{\mathrm{area}})$ values is observed. For the majority of $\ln([\mathrm{CH_4}]_{\mathrm{area}})$ distributions the QQ plots suggest normality, confirming the evaluation of the test statistics. In some cases, a failed test statistic may be due to the presence of outliers, while the QQ plot for the remaining distribution suggests normality. Notwithstanding, severe departures from normality exist.

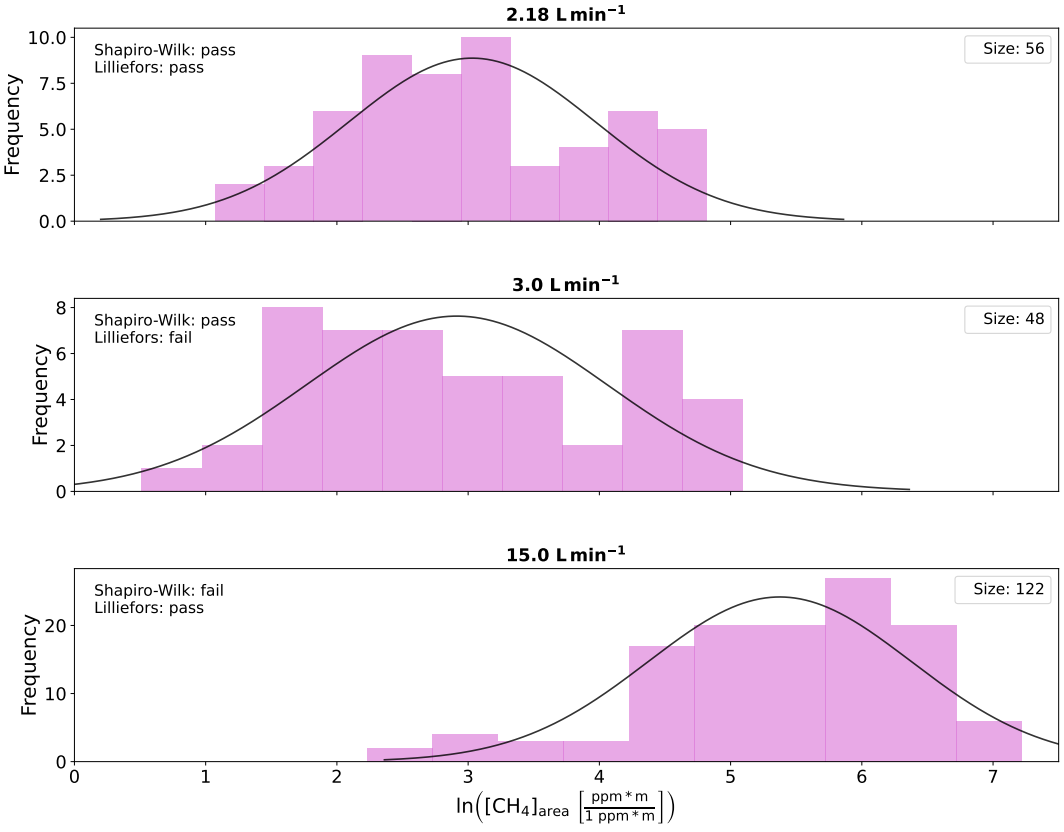


(a) Histogram

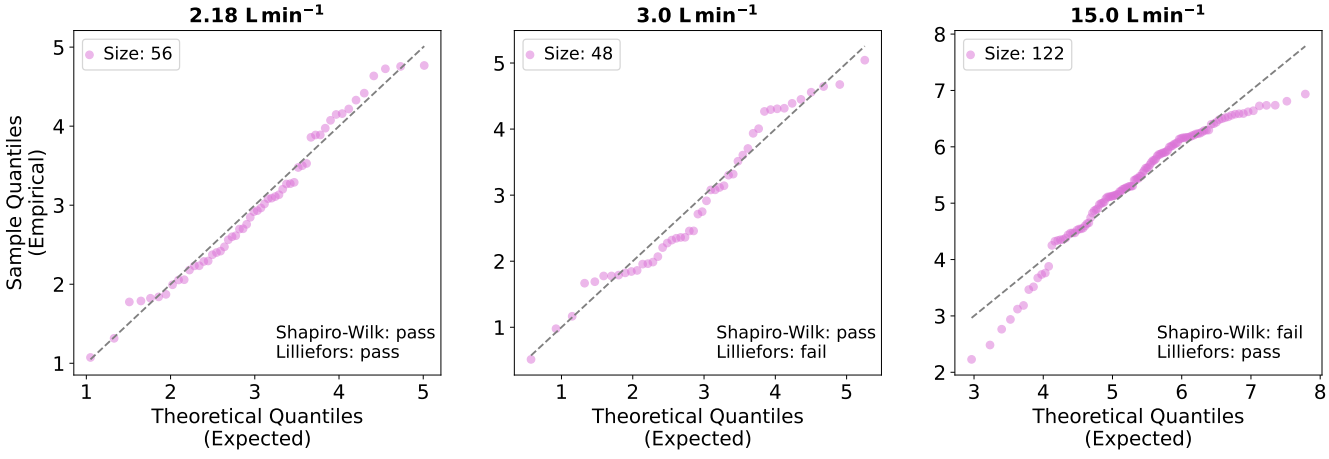


(b) Quantile-Quantile plot

Figure S16. ~~Rotterdam~~**Rotterdam**: Assessment of log-normality. (a) Histogram of the logarithmically transformed integrated peak area of the measured CH_4 enhancements. Each histogram represents areas measured at a given release rate. A Gaussian distribution is fitted to the data and the results of two normality tests (1: Shapiro-Wilk and 2: Lilliefors) are shown as well as the size of the data set. (b) Quantile-Quantile plot of the logarithmically transformed integrated peak area ($\ln([CH_4]_{area})$) versus a normal distribution for each release rate separately.

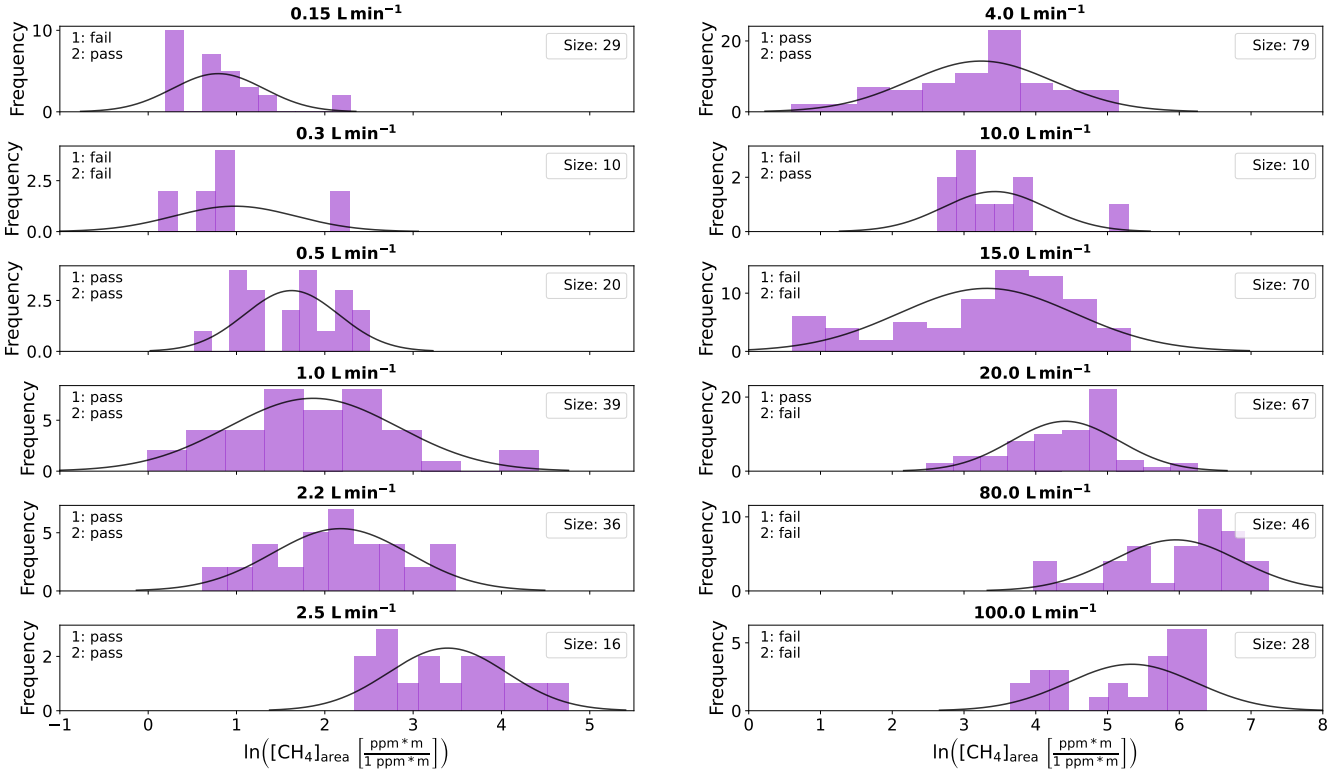


(a) Histogram

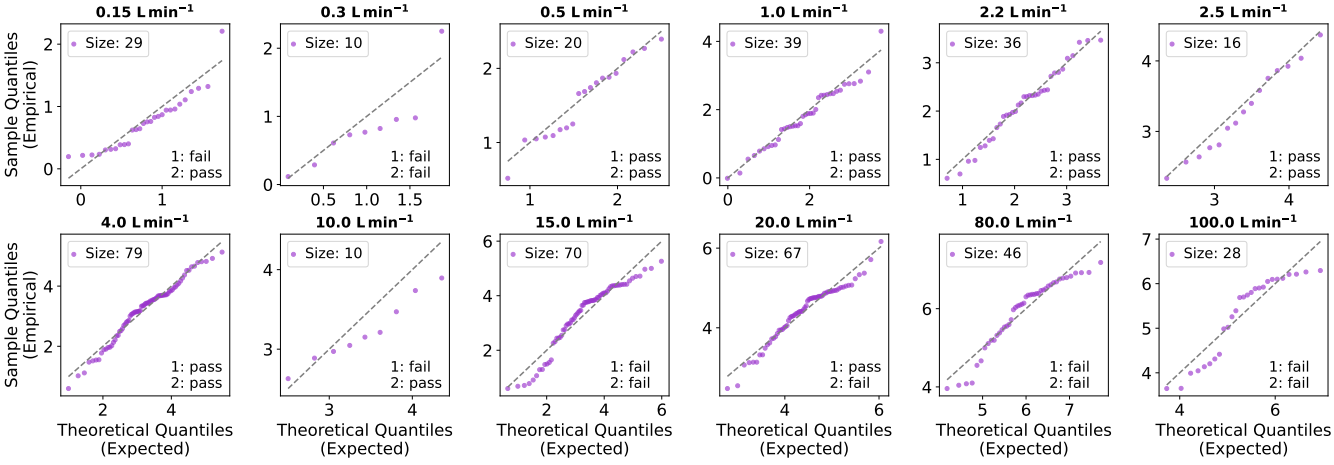


(b) Quantile-Quantile plot

Figure S17. ~~Utrecht~~**Utrecht I**: Assessment of log-normality. (a) Histogram of the logarithmically transformed integrated peak area of the measured CH_4 enhancements. Each histogram represents areas measured at a given release rate. A Gaussian distribution is fitted to the data and the results of two normality tests (Shapiro-Wilk and Lilliefors) are shown as well as the size of the data set. (b) Quantile-Quantile plot of the logarithmically transformed integrated peak area ($\ln([CH_4]_{area})$) versus a normal distribution for each release rate separately.

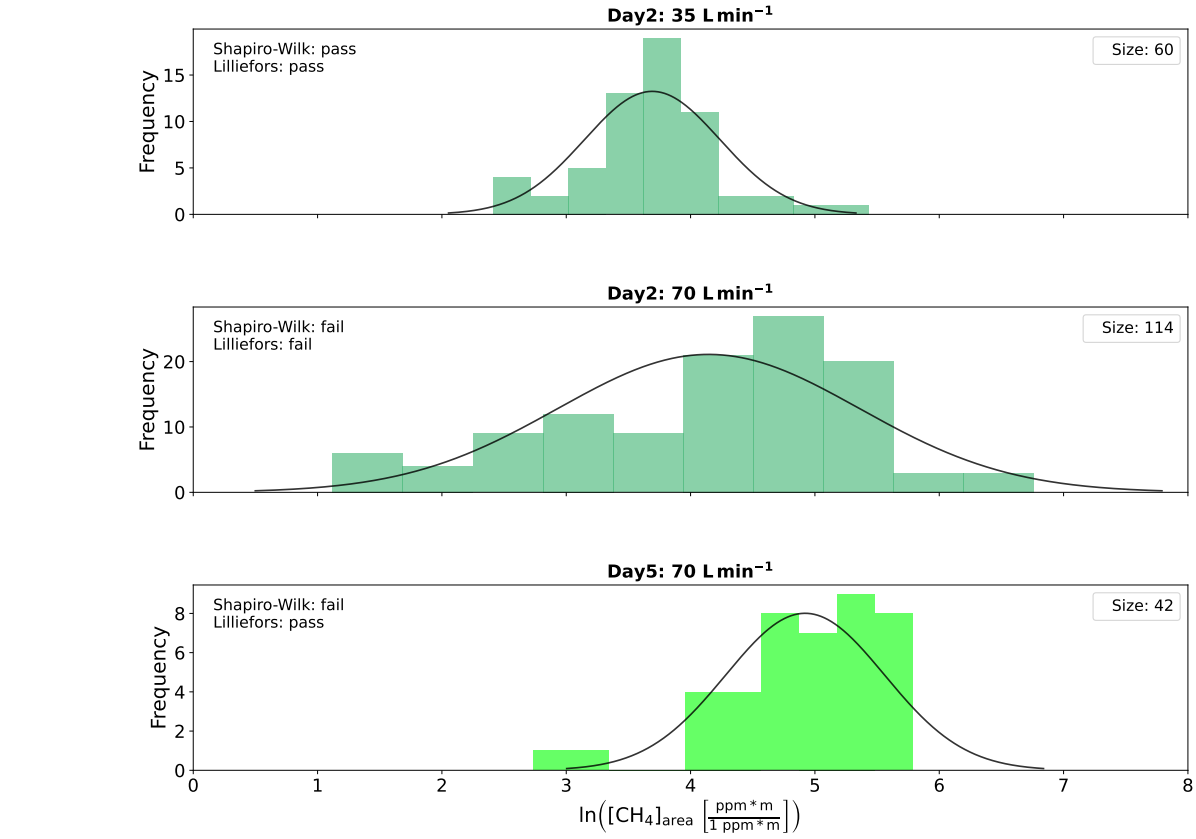


(a) Histogram

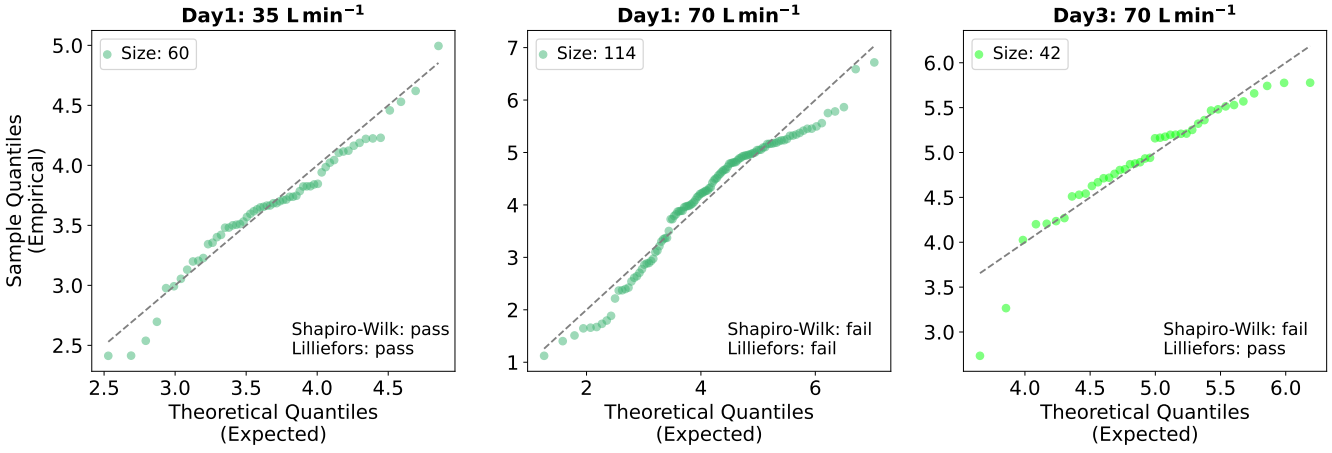


(b) Quantile-Quantile plot

Figure S18. ~~Utrecht II~~Utrecht II: Assessment of log-normality. (a) Histogram of the logarithmically transformed integrated peak area of the measured CH_4 enhancements. Each histogram represents areas measured at a given release rate. A Gaussian distribution is fitted to the data and the results of two normality tests (1: Shapiro-Wilk and 2: Lilliefors) are shown as well as the size of the data set. (b) Quantile-Quantile plot of the logarithmically transformed integrated peak area ($\ln([CH_4]_{area})$) versus a normal distribution for each release rate separately.

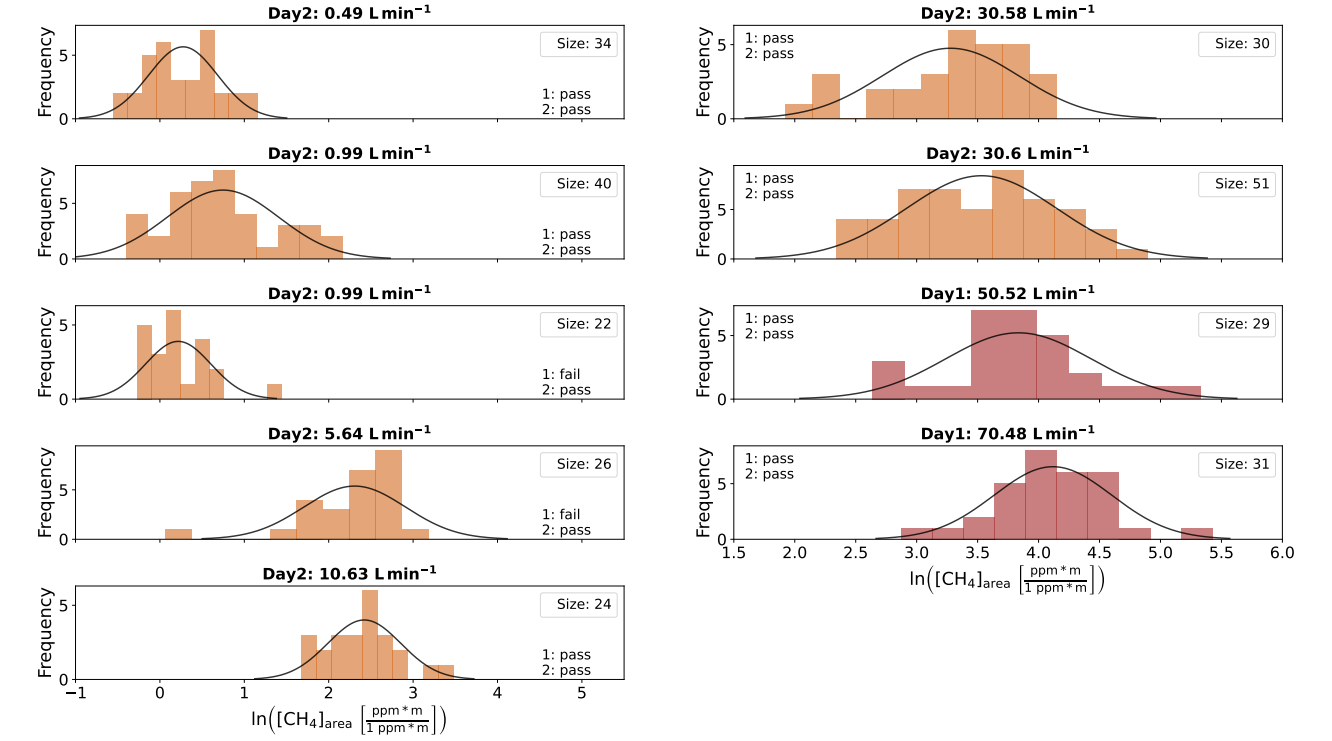


(a) Histogram

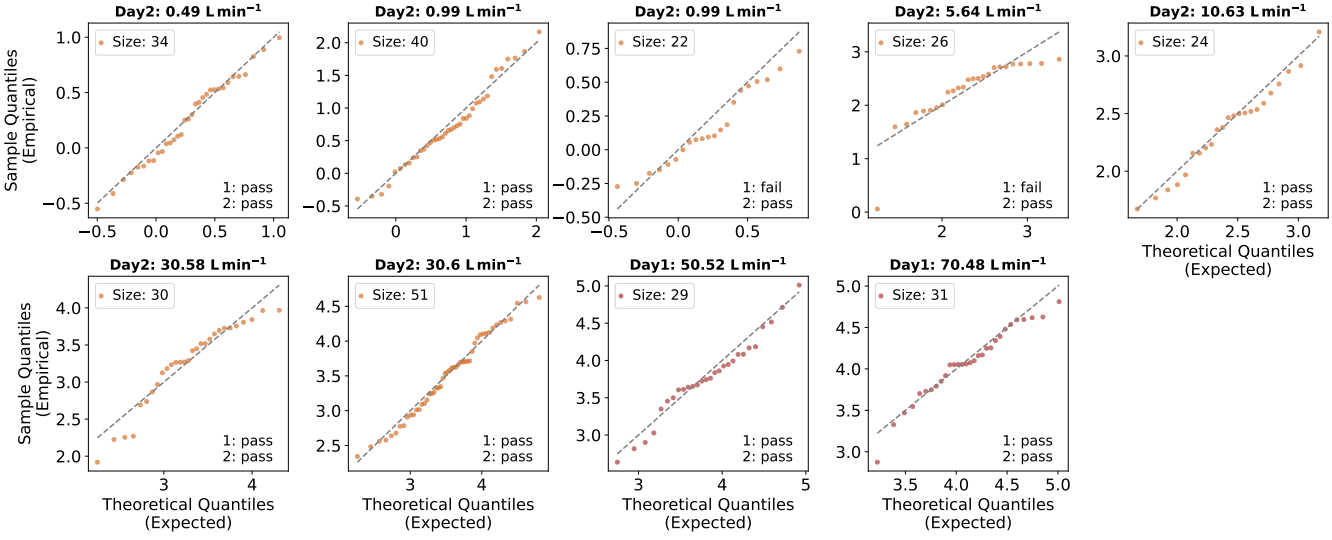


(b) Quantile-Quantile plot

Figure S19. ~~London~~**London I**: Assessment of log-normality. (a) Histogram of the logarithmically transformed integrated peak area of the measured CH_4 enhancements. Each histogram represents areas measured at a given release rate. A Gaussian distribution is fitted to the data and the results of two normality tests (Shapiro-Wilk and Lilliefors) are shown as well as the size of the data set. (b) Quantile-Quantile plot of the logarithmically transformed integrated peak area ($\ln([CH_4]_{area})$) versus a normal distribution for each release rate separately.



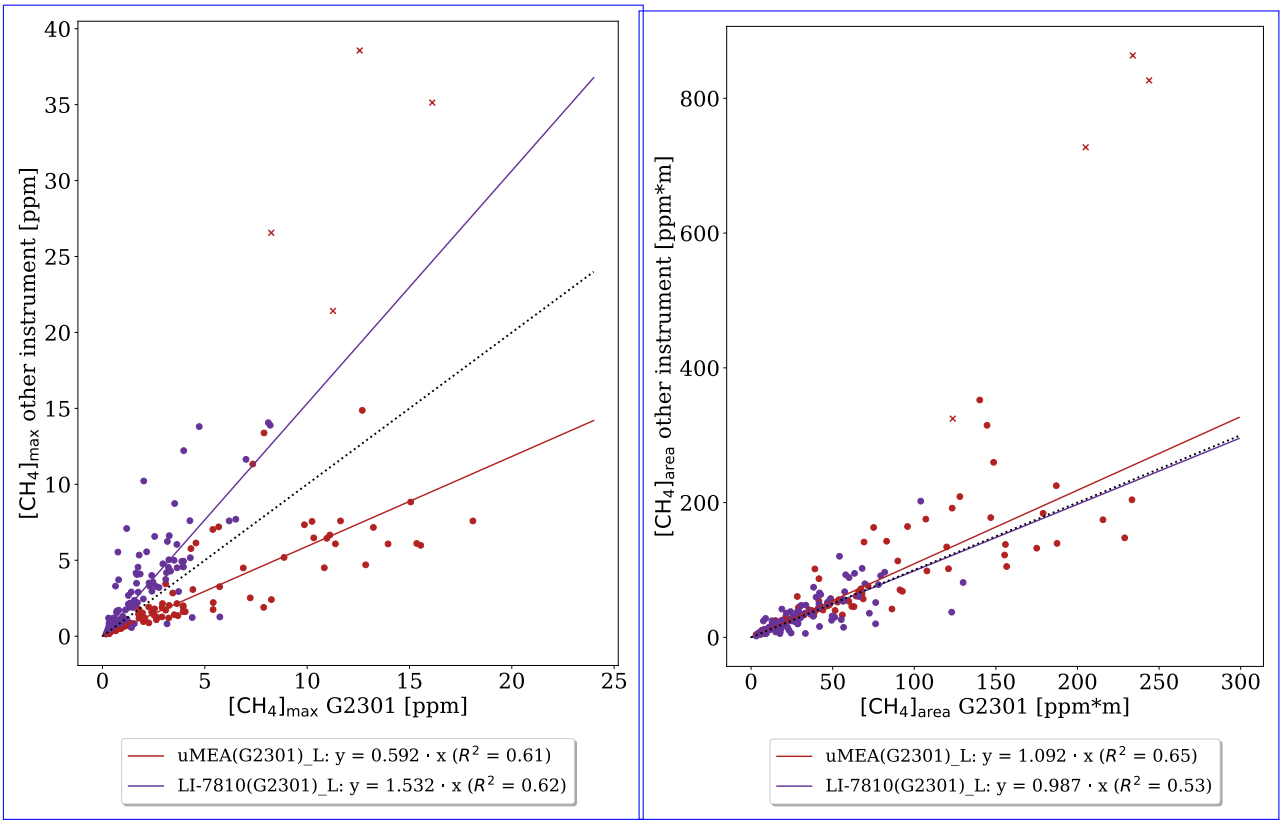
(a) Histogram



(b) Quantile-Quantile plot

Figure S20. ~~London II~~**London II**: Assessment of log-normality. (a) Histogram of the logarithmically transformed integrated peak area of the measured CH_4 enhancements. Each histogram represents areas measured at a given release rate. A Gaussian distribution is fitted to the data and the results of two normality tests (1: Shapiro-Wilk and 2: Lilliefors) are shown as well as the size of the data set. (b) Quantile-Quantile plot of the logarithmically transformed integrated peak area ($\ln([CH_4]_{area})$) versus a normal distribution for each release rate separately.

S8 Instrument Performance: Peak Maximum and Spatial Peak Area



(a) Peak Maximum (b) Spatial Peak Area

Figure S21. Comparison of peak maximum (a) and spatial peak area (b) from different instruments in London I (Day1 and Day2). Regression fits with intercept 0 are applied to the data for each instrument. The results from the uMEA and LI-7810 analyzers are plotted on the y-axis and the results from the G2301-m instrument on the x-axis. The black dotted line represents the 1:1 line. (Peaks exceeding a maximum of 20 ppm are marked with an 'x' and were excluded from the fitting process.)

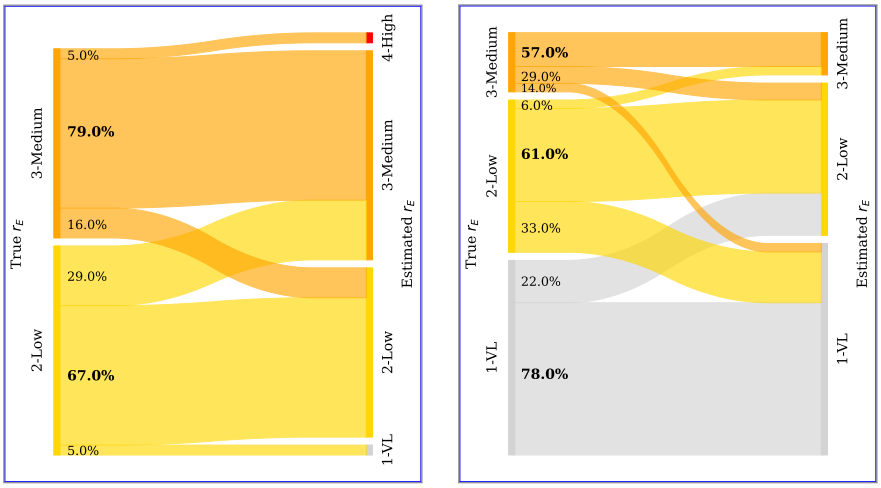
S9 Categorization of Emission Rate per Location

For each peak, the corresponding emission rate was estimated using the empirical function derived from the total dataset as presented in the main manuscript. Subsequently, a category was assigned to each peak, depending on the estimated emission size. In Tab. S8 the four different categories (1-Very low, 2-Low, 3-Medium and 4-High) are defined as well as corresponding maxima ~~ranges~~ and area ranges for the two emission rate estimation methods, e.g. a peak with a spatial peak area of 56 $\mathrm{ppm} * \mathrm{m}$ ($25 < 56 > 109\ \mathrm{ppm} * \mathrm{m}$) will be assigned an emission rate between 6-40 $\mathrm{L min}^{-1}$ and therefore categorized as a medium leak. As measurements in different locations can exhibit different offsets in their distribution, the categorization performance varies

across locations. This is visualized in Fig. S22–Fig. S23and . . This means that at certain locations, the statistical model may not perform well due to specific characteristics of the built environment. For example, narrow streets with tall buildings can either create tunnelling effects with high wind velocities or block the wind, resulting in very low velocities, depending on their orientation relative to the main wind direction and surrounding structures. Even at the same location, varying weather conditions on different days can influence the plume shape, leading to fluctuations in categorization success rates. For instance, while on Day1 of the London II controlled release experiment over 50% of low-emission-rate peaks were correctly classified, only 20% of peaks in the same emission category were correctly classified the following day.

Table S8. Natural gas distribution network CH_4 emission categories. Corresponding maxima ranges and area ranges are given for the two emission rate estimation methods.

Class	Emission Rate $[\mathrm{Lmin}^{-1}]$	Weller eq. [ppm]	Area eq. [ppm ∗ m]
High	> 40	> 7.6	> 109
Medium	$6 - 40$	$1.6 - 7.6$	$25 - 109$
Low	$0.5 - 6$	$0.2 - 1.59$	$3.7 - 25$
Very Low	< 0.5	< 0.2	< 3.7



(a) Toronto Day1 (b) Toronto Day2

Figure S22. Categorization performance for data obtained in Toronto. The left y axis represents the true emission rate r_E, where the width of the bars indicate the amount of plumes belonging to each emission category (categories: 1-Very Low, 2-Low, 3-Medium and 4-High). The right y axis represents the categories estimated by the statistical model and the connecting lines visualize the amount of plumes from each category pool which the algorithm classifies into another (or the same) category.

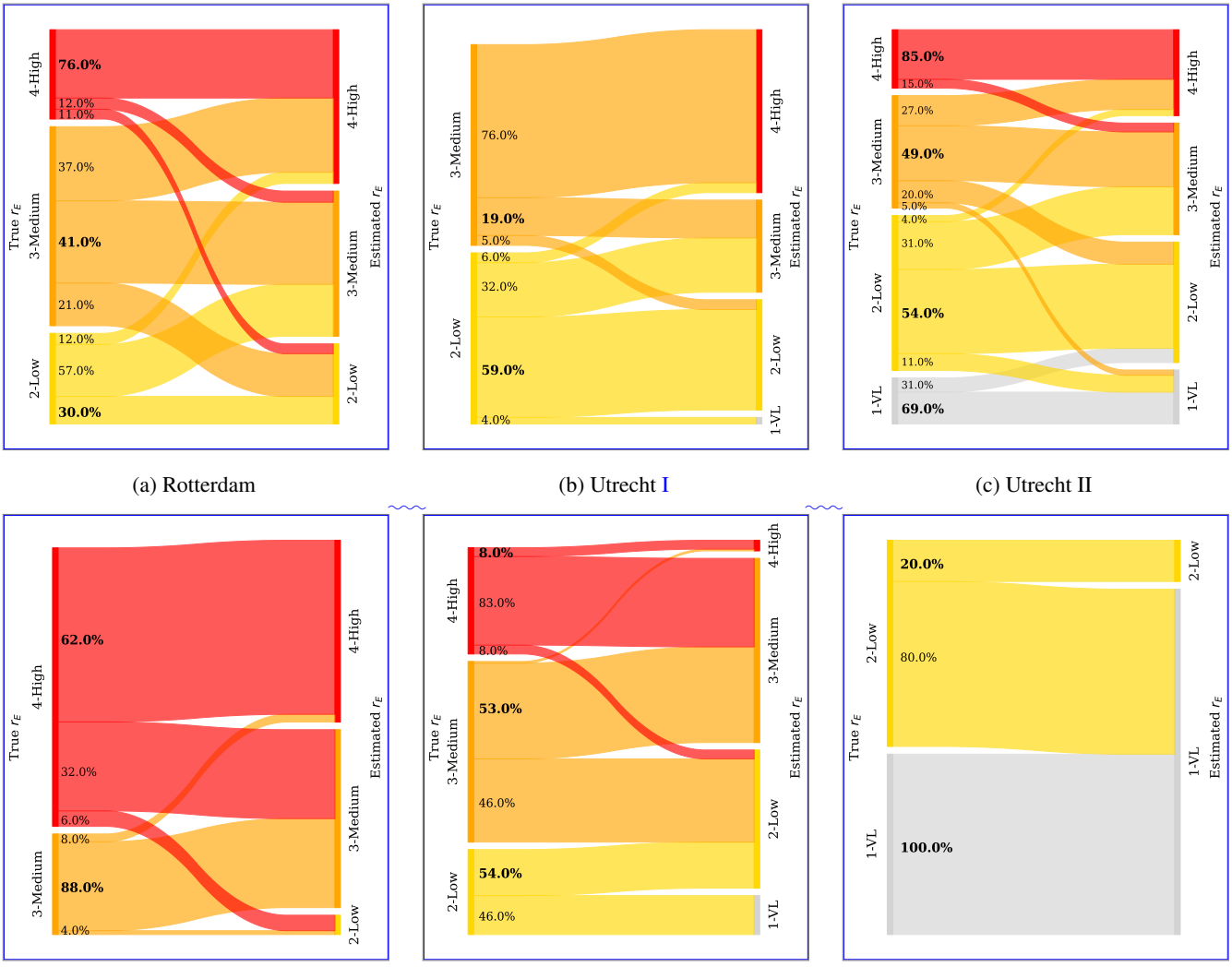


Figure S23. Categorization performance for data obtained in Rotterdam, Utrecht I, Utrecht II, London I and London II. The left y axis represents the true emission rate r_E, where the width of the bars indicate the amount of plumes belonging to each emission category (categories: 1-Very Low, 2-Low, 3-Medium and 4-High). The right y axis represents the categories estimated by the statistical model and the connecting lines visualize the amount of plumes from each category pool which the algorithm classifies into another (or the same) category.

S10 Influence of Sampling Effort

S10.1 Hypothetical Distributions

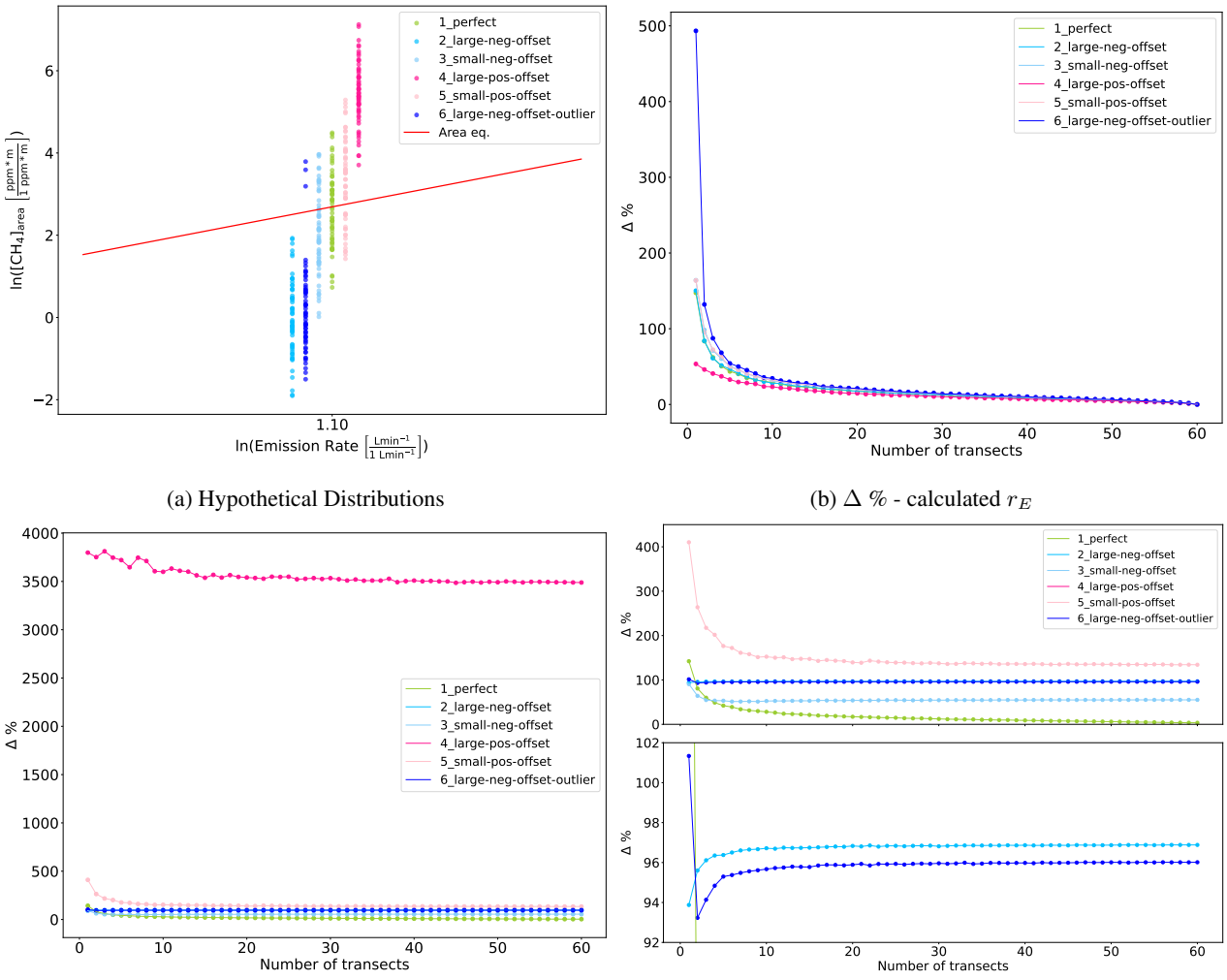
In order to illustrate the behaviour of sampling multiple times at the same locations, we present results for some selected hypothetical distributions with means falling above or below the empirical equation derived from the totality of all measurements in the main manuscript. Fig. S24 and Fig. S25 display hypothetical distributions randomly sampled with standard deviations of 1 and different offsets for the release rates 3 Lmin^{-1} and 50 Lmin^{-1}. A 'perfect' distribution is included for which the mean corresponds to the $\ln([\mathrm{CH_4}]_{\mathrm{area}})$ value that we expect for this release rate following the Area eq. The offsets are selected so that, in log space, they maintain an equal distance from the mean of the ideal distribution in both positive and negative directions (e.g., $\pm$ 0.7 for the distributions with a small positive and small negative offsets).

For the 3 Lmin^{-1} case, the percentage difference of the estimated release rates to the calculated mean emission rate decreases to 0 as expected (Fig. S24b). The absolute percentage error decreases for a higher number of transects for all distributions except for the ones with a large negative offset. The percentage error is greater for the distribution with a small positive offset compared to that with a small negative offset. The distribution with the large positive offset shows the highest percentage error relative to the true emission rate. Interestingly, however, it has the smallest percentage error when compared to the calculated mean emission rate.

For the 50 Lmin^{-1} case, the percentage difference of the estimated release rates to the calculated mean emission rate decreases to 0 as expected, except for the distribution with the high positive offset (Fig. S25b). This is due to the fact that the mean of the distribution corresponds to a release rate higher then the cap of 200 Lmin^{-1}. The absolute percentage error decreases only for the perfect distribution and the ones with a small negative or positive offset.

The two example shows that generally the error in estimations decreases when including more transects. However, in case of a large deviation of the measurement distribution from the one we expect following our method the behaviour can differ. Apart from the offset itself, other parameters play a role such as the imposed emission rate cap and likely the standard deviation or presence of outlier.

Fig. S26 and Fig. S27 illustrate the calculations steps from the $\ln([\mathrm{CH_4}]_{\mathrm{area}})$ distribution to the final mean absolute percentage differences. In panel (a), the underlying distribution is shown in gray, with black markers representing the means of different Monte Carlo samples of size N. As sample size increases, the spread of the sample means narrows until it converges to the overall distribution mean at ~~N=60~~ $N = 60$, the population size. Panel (b) depicts the emission rate estimates derived from the sample means in (a). As sample size grows, variability in the emission rate estimates diminishes. Further, it is evident that larger overestimations than underestimations occur. Panel (c) presents the percentage deviations of the estimated emission rate from the true rate, showing both positive and negative directions. Finally, panel (d) displays the absolute percentage deviations, similar to Figure 5 in the main manuscript.



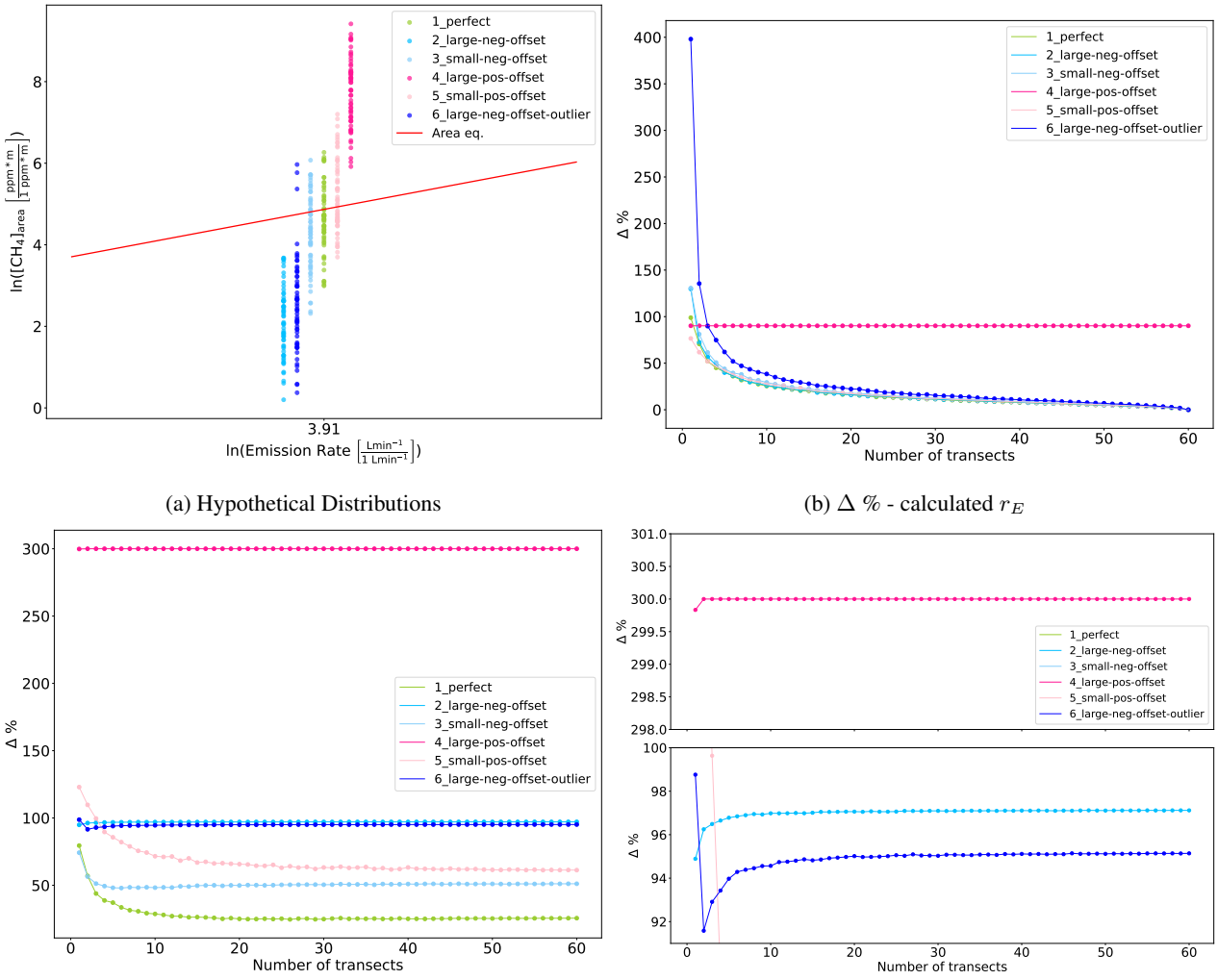
(a) Hypothetical Distributions

(b) Δ % - calculated r_E

(c) Δ % - true r_E

(d) Δ % - true r_E Zoom

Figure S24. Hypothetical distributions for a release rate 3 Lmin^{-1} with different offsets from the perfect distribution.



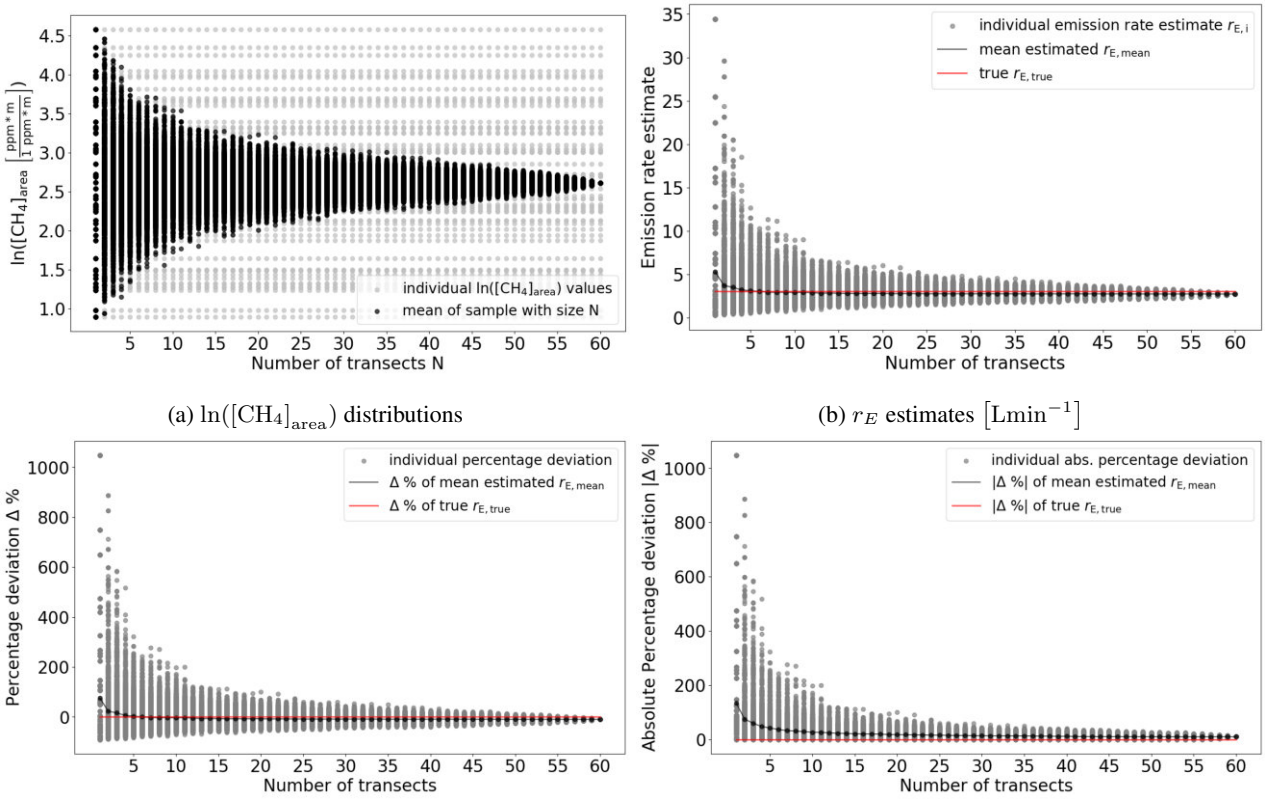
(a) Hypothetical Distributions

(b) Δ % - calculated r_E

(c) Δ % - true r_E

(d) Δ % - true r_E Zoom

Figure S25. Hypothetical distributions for a release rate 50 Lmin^{-1} with different offsets from the perfect distribution.



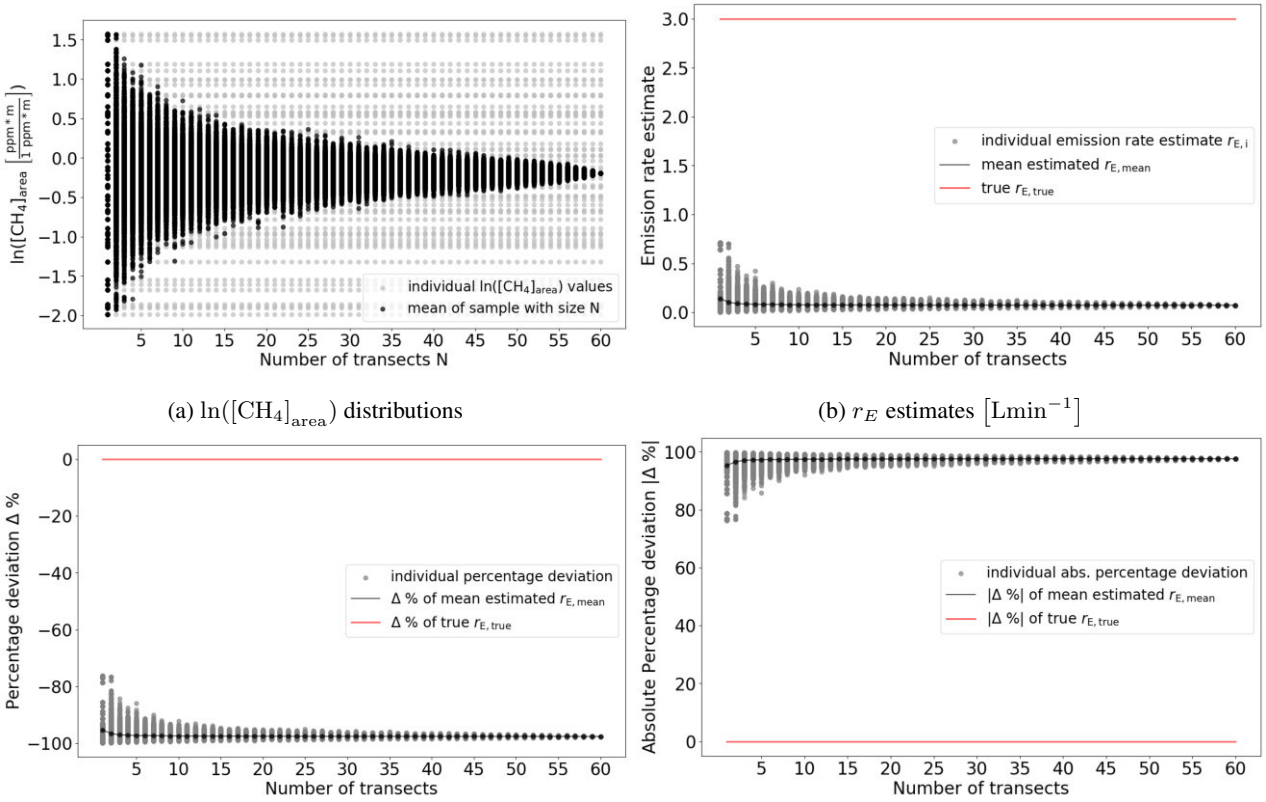
(a) $\ln([\mathrm{CH_4}]_{\mathrm{area}})$ distributions

(b) r_E estimates $\left[\mathrm{Lmin}^{-1}\right]$

(c) Δ % - r_E from true r_E

(d) Absolute Δ % - r_E from true r_E

Figure S26. Perfect distribution: Visualization of different calculation steps in the analysis of the benefit of multiple transects.



(a) $\ln([\mathrm{CH_4}]_{\mathrm{area}})$ distributions

(b) r_E estimates $\left[\mathrm{Lmin}^{-1}\right]$

(c) Δ % - r_E from true r_E

(d) Absolute Δ % - r_E from true r_E

Figure S27. Distribution with large negative offset: Visualization of different calculation steps in the analysis of the benefit of multiple transects.

Code and data availability. The python code and a sub-sample of the data used to produce the results in this article are available on GitHub:
240 https://github.com/judith-tettenborn/CRE_CH4Quantification.git

References

Ars, S., Vogel, F., Arrowsmith, C., Heerah, S., Knuckey, E., Lavoie, J., Lee, C., Pak, N. M., Phillips, J. L., and Wunch, D.: Investigation of the Spatial Distribution of Methane Sources in the Greater Toronto Area Using Mobile Gas Monitoring Systems, Environmental Science & Technology, 54, 15 671–15 679, https://doi.org/10.1021/acs.est.0c05386, 2020.

Biecek, P. and Burzykowski, T.: Explanatory Model Analysis: Explore, Explain, and Examine Predictive Models, CRC Press, New York, ISBN 978-0-367-13559-1, https://www.taylorfrancis.com/books/mono/10.1201/9780429027192/explanatory-model-analysis-przemyslaw-biecek-tomasz-burzykowski, 2021.

Flatt, C. and Jacobs, R. L.: Principle Assumptions of Regression Analysis: Testing, Techniques, and Statistical Reporting of Imperfect Data Sets, Advances in Developing Human Resources, 21, 484–502, https://doi.org/10.1177/1523422319869915, 2019.

Keskin, S.: Comparison of Several Univariate Normality Tests Regarding Type I Error Rate and Power of the Test in Simulation Based Small Samples, Journal of Applied Science Research, 2, 296–300, https://www.academia.edu/download/88717042/296-300.pdf, 2006.

Nimon, K.: Statistical Assumptions of Substantive Analyses Across the General Linear Model: A Mini-Review, Frontiers in Psychology, 3, https://doi.org/10.3389/fpsyg.2012.00322, 2012.

Razali, N. M. and Wah, Y. B.: Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests, Journal of Statistical Modeling and Analytics, 2, 21–33, 978-967-363-157-5, 2011.

von Fischer, J. C., Cooley, D., Chamberlain, S., Gaylord, A., Griebenow, C. J., Hamburg, S. P., Salo, J., Schumacher, R., Theobald, D., and Ham, J.: Rapid, Vehicle-Based Identification of Location and Magnitude of Urban Natural Gas Pipeline Leaks, Environmental Science & Technology, 51, 4091–4099, https://doi.org/10.1021/acs.est.6b06095, 2017.

Von Storch, H. and Zwiers, F. W.: Statistical Analysis in Climate Research, Cambridge University Press, ISBN 978-1-139-42509-4, 2002.

Weller, Z. D., Roscioli, J. R., Daube, W. C., Lamb, B. K., Ferrara, T. W., Brewer, P. E., and von Fischer, J. C.: Vehicle-Based Methane Surveys for Finding Natural Gas Leaks and Estimating Their Size: Validation and Uncertainty, Environmental Science & Technology, 52, 11 922–11 930, https://doi.org/10.1021/acs.est.8b03135, 2018.