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

The residence time was determined on the basis of the cell temperature T_cell $[\mathrm{K}]$, cell pressure p_cell $[\mathrm{Pa}]$, cell volume V_cell $\left[\mathrm{m}^3\right]$ and flow rate Q_cell $[\mathrm{slm}]$ specified by the manufacturers. The units in brackets specify the units in which the different quantities have to be inserted into the equation.

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

$$V_\mathrm{norm} = \frac{p_\mathrm{cell} \cdot V_\mathrm{cell} \cdot R \cdot T_\mathrm{norm}}{R \cdot T_\mathrm{cell} \cdot p_\mathrm{norm}} \tag{1}$$

Then, given the flow rate, the residence time was determined from V_norm and the flow rate Q_cell as:

$$\tau = \frac{V_\mathrm{norm}}{Q_\mathrm{cell}} \tag{2}$$

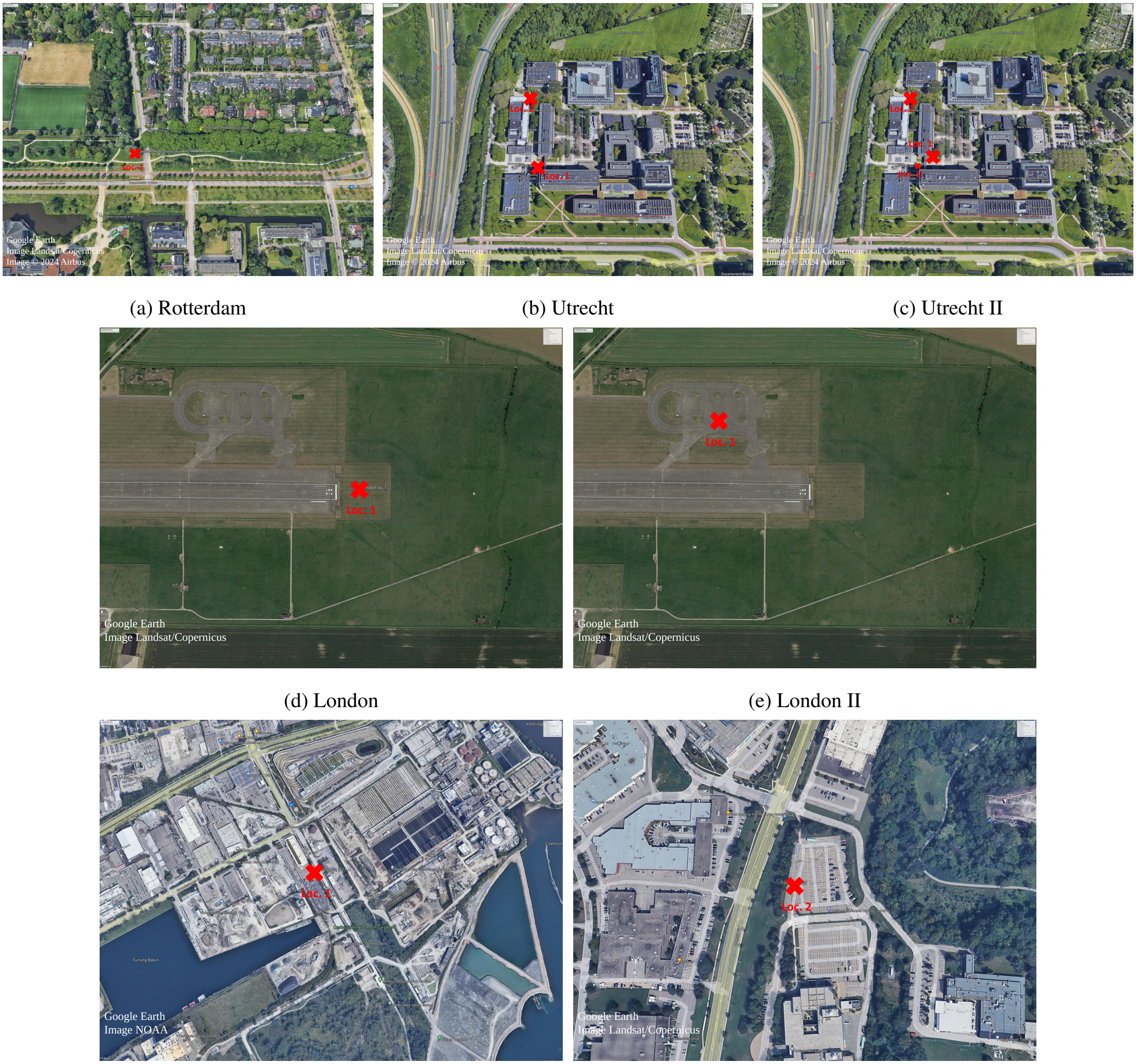
Table S1. Overview of instrument characteristics of analyzers deployed in the controlled release experiments. In cases where flow rate varied, the bold numbers were used for the calculation of the residence time.

GHG Analyzer	T_{cell} [°C]	p_{cell} [mbar]	V_{cell} [mL]	Air Volume in Cell V(p, T) [mL]	Q_{cell} [slm]	τ [s]	Measurement Frequency [Hz]
G2301[a]	45	190	50	8.8	0.4	1.3	0.36
G4302[a]	35	600	35	21.8	2.4	0.5	1
G2401[a]	35	186	35	6.0	0.4	0.9	0.4
uMEA[b]	.	.	.	345	.	.	1
UGGA[b]	25	186	345	63.3	**2**-4	1.9	0.1-1
LI-7810[c]	55	390	6.41	2.2	0.25	0.5	1
TILDAS[d]	25	40	500	19.7	6	0.2	2
Mira Ultra[e]	42	240	60	13.5	0.3-**0.6**	1.3	1
MGA10[f]	27	80	500	39.2	4	0.6	2

[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.

S2 Description Controlled Release Experiments

S2.1 Release Locations



(a) Rotterdam

(b) Utrecht

(c) Utrecht II

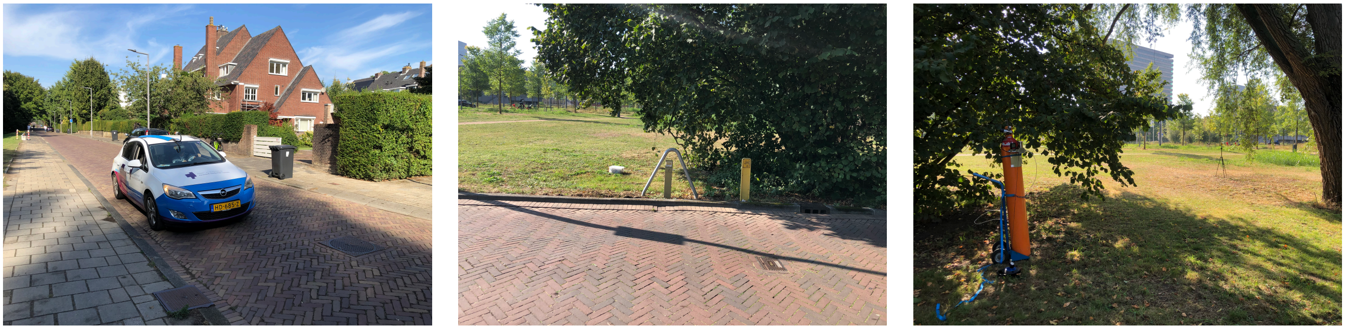
(d) London

(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.

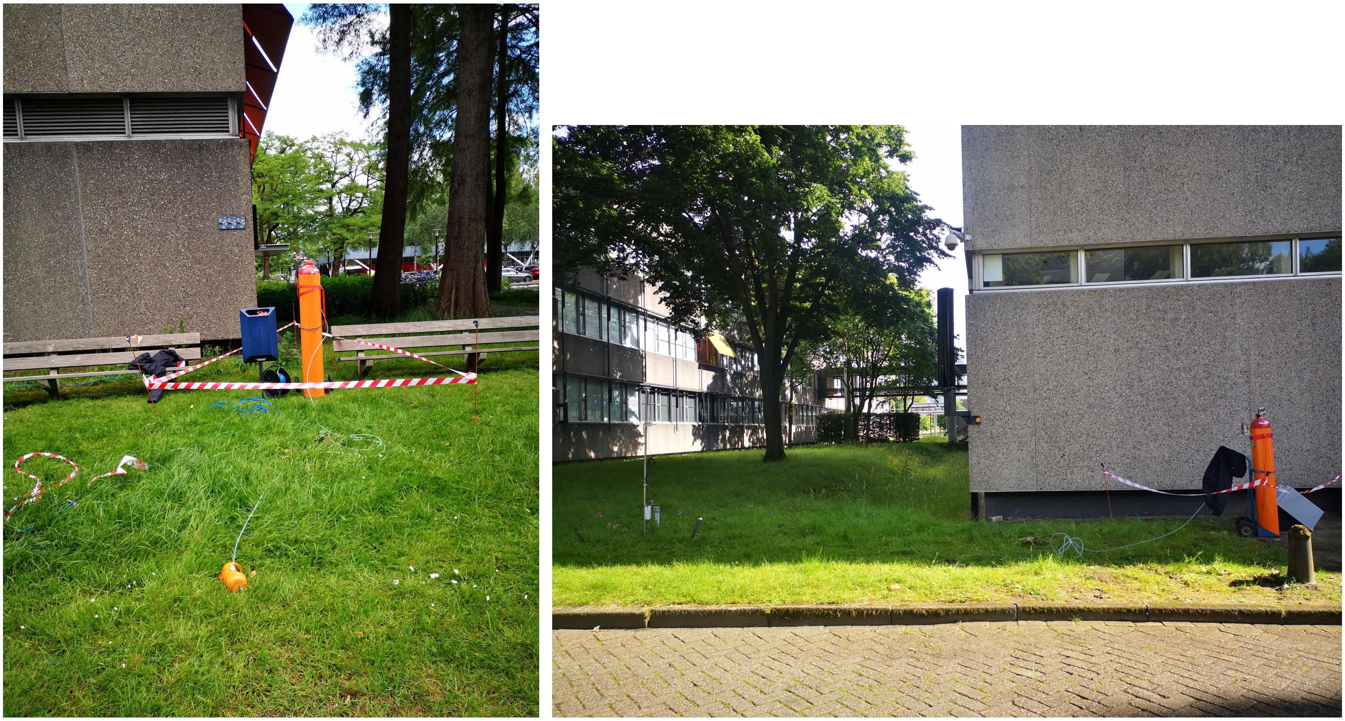


(a) UUAQ car

(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

Rotterdam

The control range of the Alicat mass flow controller is 0-100 $\mathrm{L\,min^{-1}}$ under standard conditions with a measurement accuracy of ±(0.8% of reading + 0.2% of full scale). CH_4 gas was released via a 1/4' O.D. teflon tube from the surface level at a distance of about 1-3 m from the street. For the higher flow rates (above 40 $\mathrm{Lmin^{-1}}$) the inlet line was moved about 5 m away from the street.

The 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 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 Mira Ultra instrument has a measurement frequency of 1 Hz, a sensitivity of < 2 $\mathrm{ppbs^{-1}}$ 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.

Two instruments, a Miro MGA10 analyzer and Aerodyne TILDAS Dual Laser Trace Gas Analyzer were operated in the measurement trailer of a truck operated by TNO. The MGA10 measured at 1 Hz with precision of 1 ppb within the measurement range 0-200 ppm. The TILDAS analyzer measured at 1 Hz with precision of 2.4 ppb and had a response time equal to about 2 s.

Utrecht

Utrecht I

CH_4 was released simultaneously from two cylinders 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 rate which was controlled by the pressure reducer of the cylinder (3 different release rates spanning from 2.18 - 15 $\mathrm{Lmin^{-1}}$). CH_4 mole fraction were measured by the G2301 and G4302 devices from Picarro Inc., 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. Each complete circle took approximately 1.5 to 2 minutes.

Utrecht II

Initially, the same two release locations from the previous experiment were used. However, after encountering power supply issues with the battery powering the flow controller at location 1, the release point was moved to the opposite side of the street. At this first location, only one release rate (4 $\mathrm{Lmin^{-1}}$ for approximately 30 minutes) was applied.

Two mass flow controllers were used: an Alicat device for higher release rates and an MKS (PR 4000) controller for lower rates. At the start, the Alicat MFC was installed at location 1, and the MKS MFC at location 2. Midway through the experiment, the controllers were switched to allow the full range of release rates at both locations.

CH_4 mole fractions were measured using the G2301 and Mira Ultra devices from the IMAU van.

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 Absorption Spectroscopy (OF-CEAS) to detect gases in air. It can measure CH_4 within the range 0-100 ppm with a precision (1 σ) of 0.6 (0.25) ppb at 2 ppm with 1 (5) s averaging. Its Response time ($\mathrm{T}_{10} - \mathrm{T}_{90}$ from 0 to 2 ppm is ≤ 2 s. 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 G2301-m greenhouse gas (GHG) 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 driving pattern consists of multiple parallel legs that are oriented perpendicular to the estimated wind direction. These legs progressively move away from the source and then return to complete a series of passes at the same distance.

London II

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

Toronto

On 20 October 2021, where both a mobile bicycle-trailer-based laboratory (UGGA analyzer) and a vehicle based setup (G2401 analyzer) were deployed, the bicycle followed the vehicle at a distance of around 30 m through the same CH_4 plume. The G2401 analyzer has a precision of < 1 ppb for CH_4 over a 5 s integration period. 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.

S3 Raw Data Processing

The raw measurements taken by the G4302, G2301 and Mira Ultra instruments during the Rotterdam and Utrecht 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:

$$[\mathrm{CH_4}]_{\mathrm{calibrated}} = 1.03127068196 \cdot [\mathrm{CH_4}]_{\mathrm{raw}} - 0.15799666857 \tag{3}$$

G4302:

$$[\mathrm{CH_4}]_{\mathrm{calibrated}} = 1.01924906721 \cdot [\mathrm{CH_4}]_{\mathrm{raw}} - 0.05887406866 \tag{4}$$

Mira Ultra:

$$[\mathrm{CH_4}]_{\mathrm{calibrated}} = 1.01354227768 \cdot [\mathrm{CH_4}]_{\mathrm{raw}} - 0.05055326961 \tag{5}$$

$[\mathrm{CH_4}]$ refers to the CH_4 mole fraction in ppm.

S4 Overview Time Series

95 Figure S4 to Figure 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.

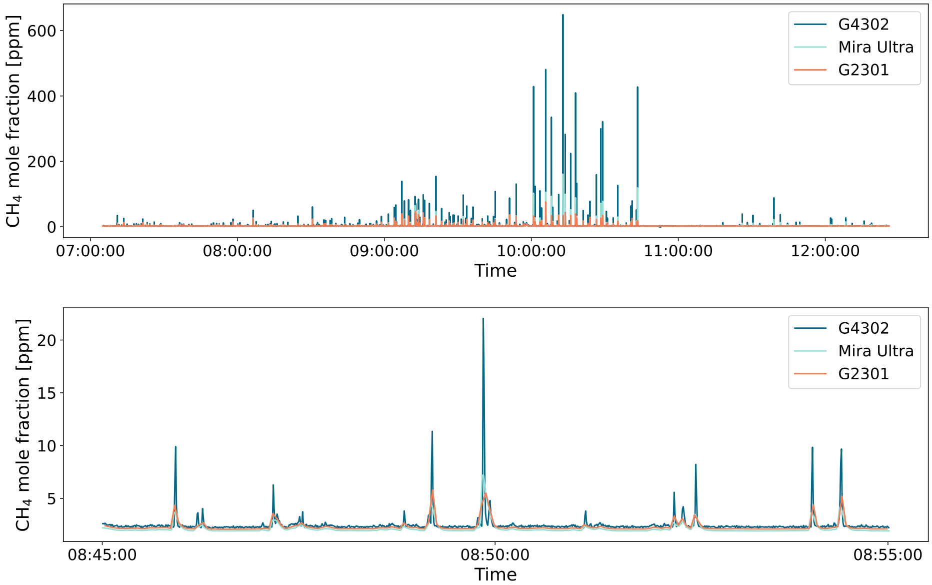


Figure S4. Rotterdam: Timeseries of CH_4 mole fraction, obtained by the G4302, G2301 and Mira Ultra device. The lower panel displays a zoom to a 10 min measurement interval. Time displayed in UTC.

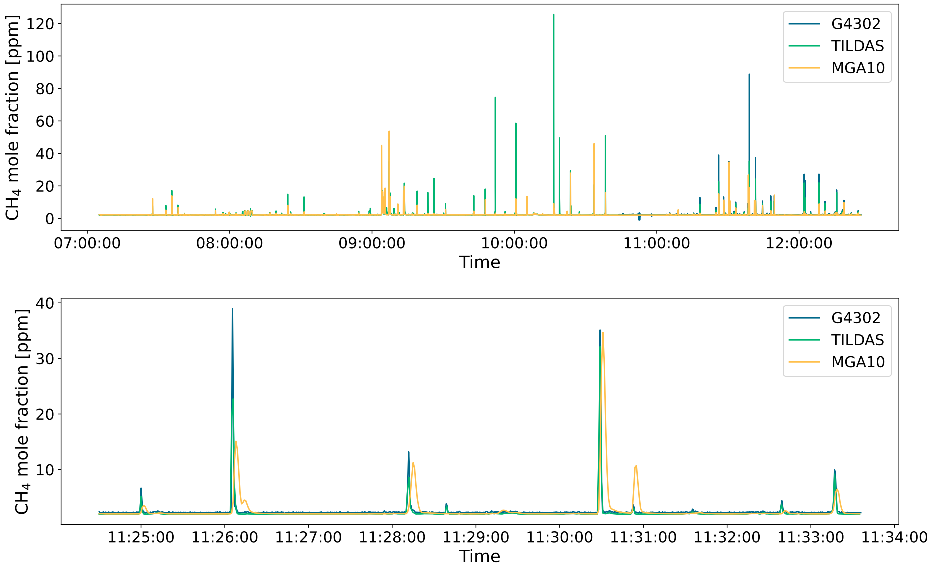


Figure S5. Rotterdam: Timeseries of CH$_4$ measurements, obtained by the MGA10, TILDAS and G4302 devices. The lower panel displays a zoom to a 10 minutes measurement interval. Time displayed in UTC.

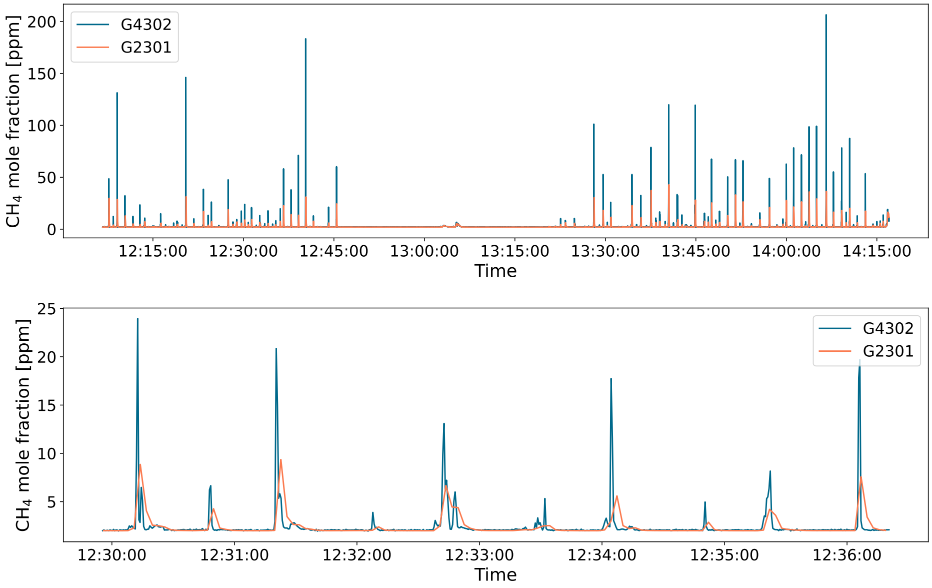


Figure S6. Utrecht: Timeseries of CH$_4$ measurements, obtained by the G4302 and G2301 devices. The lower panel displays a zoom to a 6 minutes measurement interval. Time displayed in UTC.

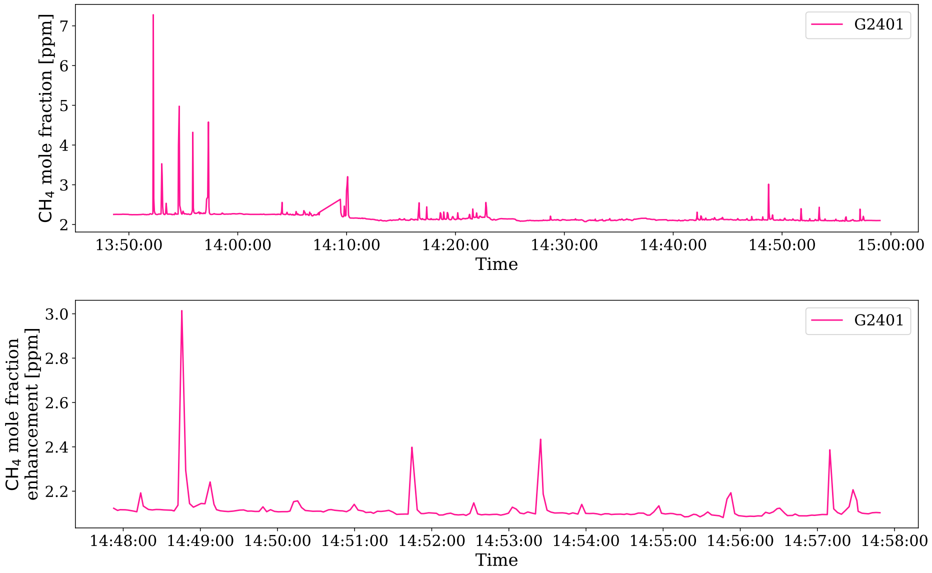


Figure S7. Toronto Day 2 - car: Timeseries of CH$_4$ measurements, obtained by the G2401 device. The lower panel displays a zoom to a 10 minutes measurement interval. Time displayed in UTC.

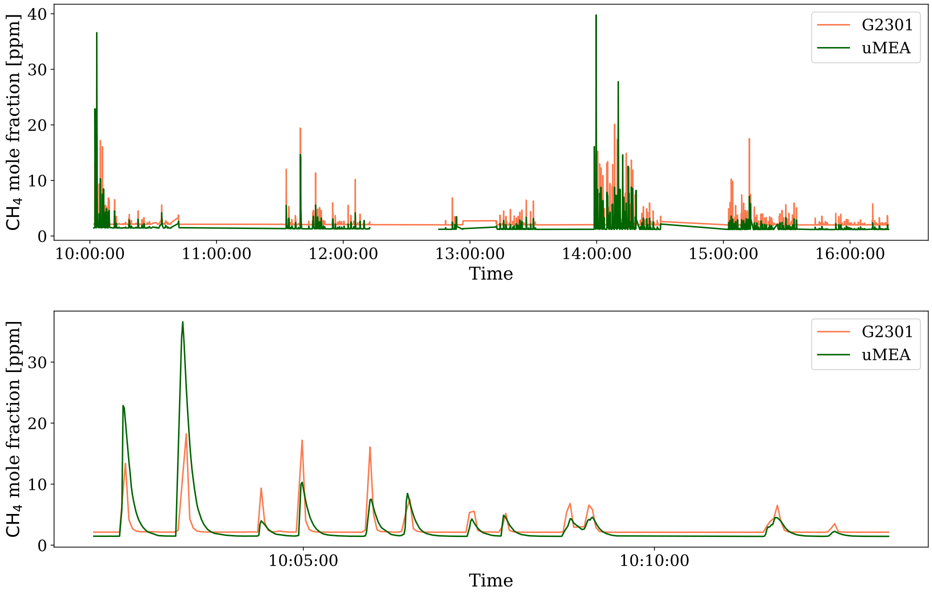


Figure S8. London Day 1: Timeseries of CH$_4$ measurements, obtained by the G2301 and uMEA devices. The lower panel displays a zoom to a 10 minutes measurement interval. Time displayed in UTC.

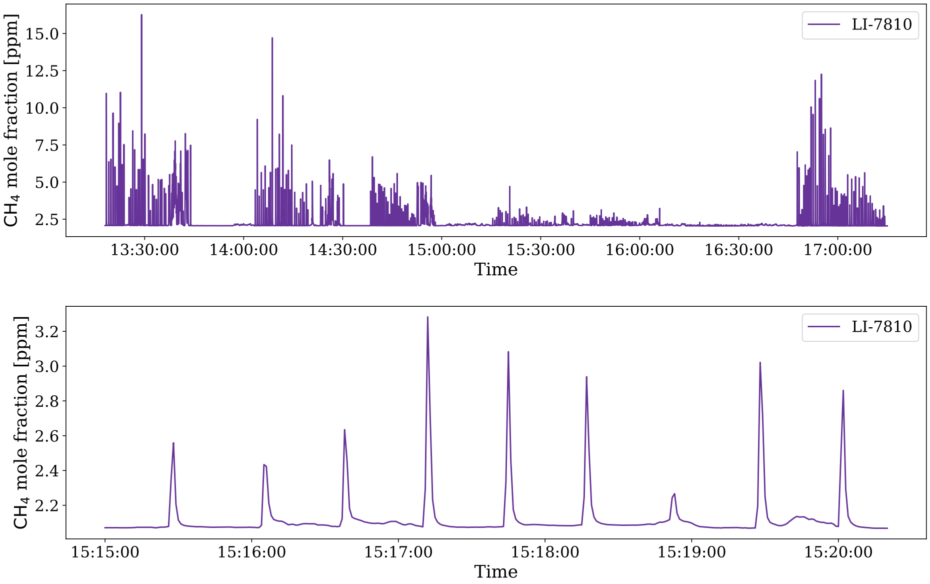
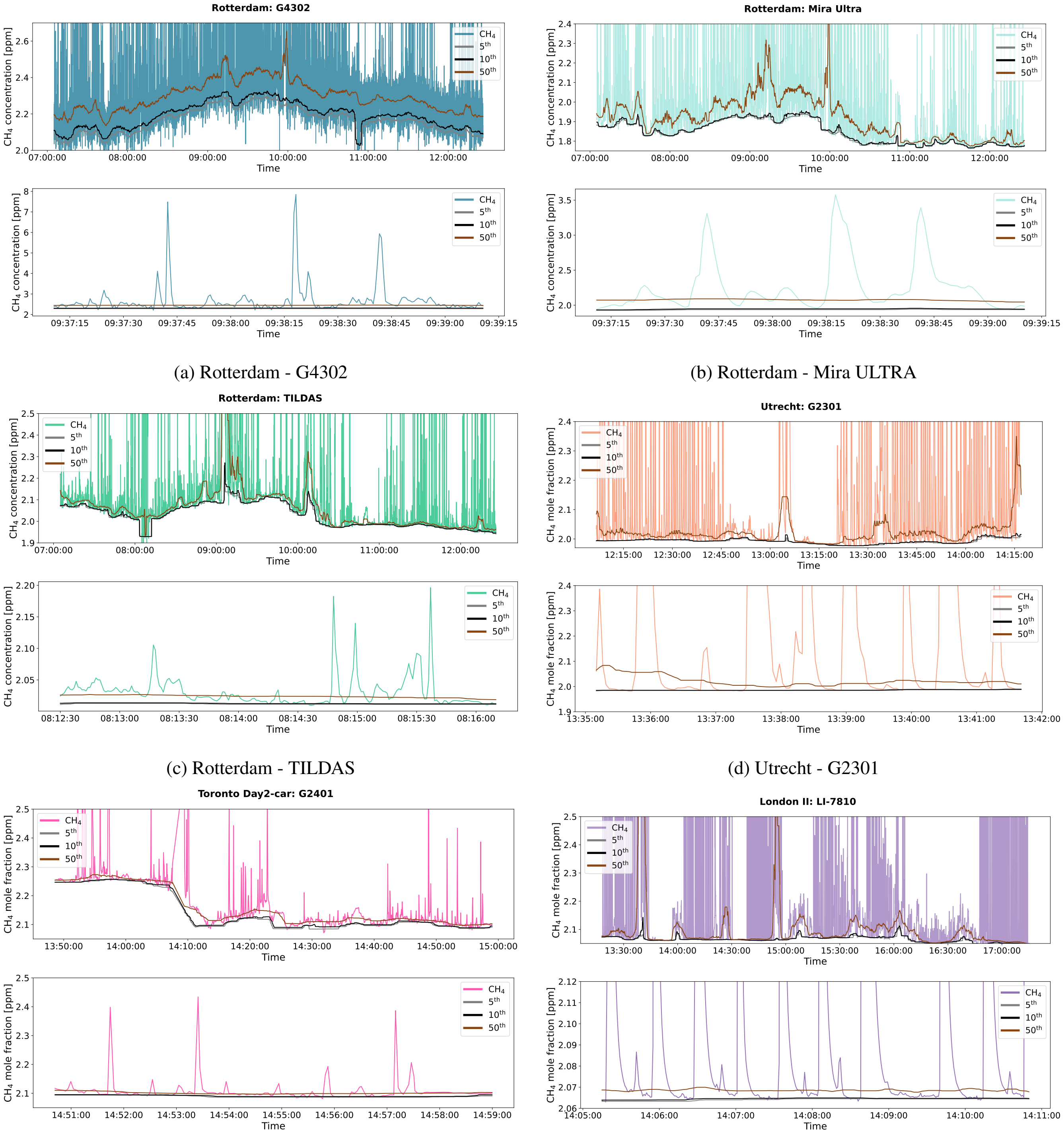


Figure S9. London II Day 1: Timeseries of CH_4 measurements, obtained by the LI-7810 device. The lower panel displays a zoom to a 5 minutes measurement interval. Time displayed in UTC.

S5 Background comparison

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)). 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) to the 50^{th} percentile (median) in Weller et al. (2018) or taking the mean in 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 (Figure S10). The 50^{th} percentile was too strongly influenced by the CH_4 release, occasionally showing up to $0.3\ \mathrm{ppm}$ higher background mole fractions compared to the 10^{th} percentile.



(a) Rotterdam - G4302

(b) Rotterdam - Mira ULTRA

(c) Rotterdam - TILDAS

(d) Utrecht - 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 spatial peak area values generally decrease with increasing distance, though the effect varies across cases and is relatively minor within a 75 m range. At some instances, 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 influence of noise. 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.

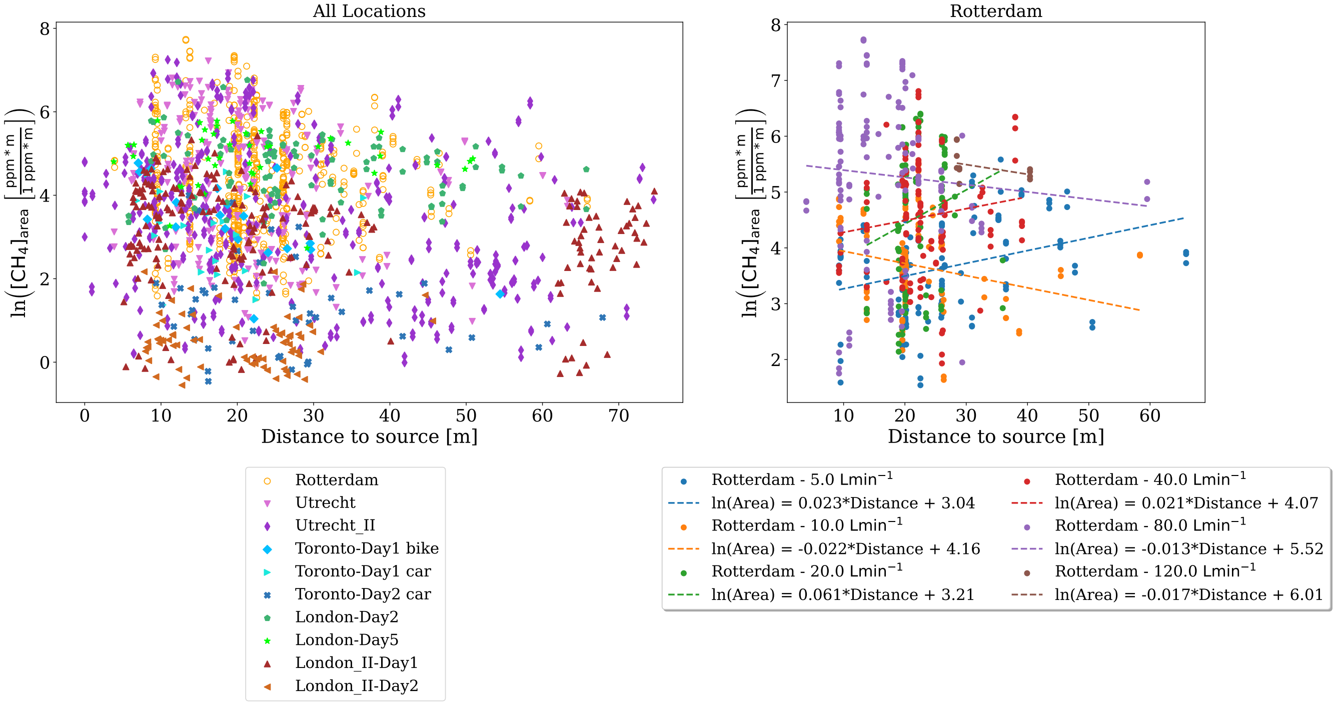


Figure S11. Logarithmic spatial peak area as function of distance to source for all individual CH_4 enhancements reported in this study (a) and Rotterdam (b). In panel (b), colours represent different release rates, with a separate linear regression fitted for each rate.

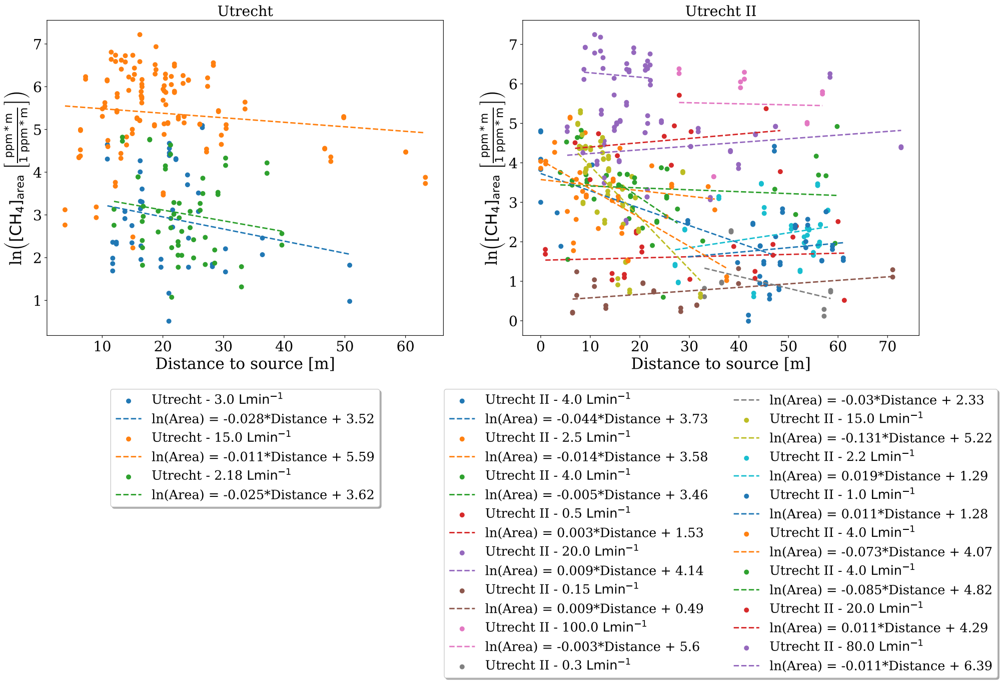


Figure S12. Logarithmic spatial peak area as function of distance to source for individual CH_4 enhancements reported in the Utrecht (a) and Utrecht II (b) controlled release experiment. Colours represent different release rates, with a separate linear regression fitted for each rate.

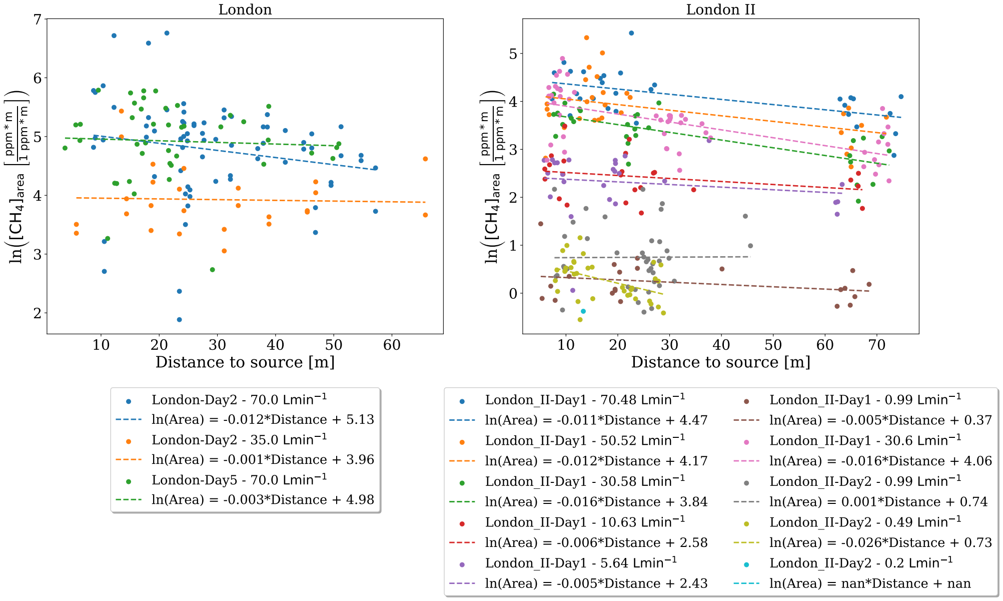


Figure S13. Logarithmic spatial peak area as function of distance to source for individual CH_4 enhancements reported in the London (a) and London II (b) controlled release experiment. Colours represent different release rates, with a separate linear regression fitted for each rate.

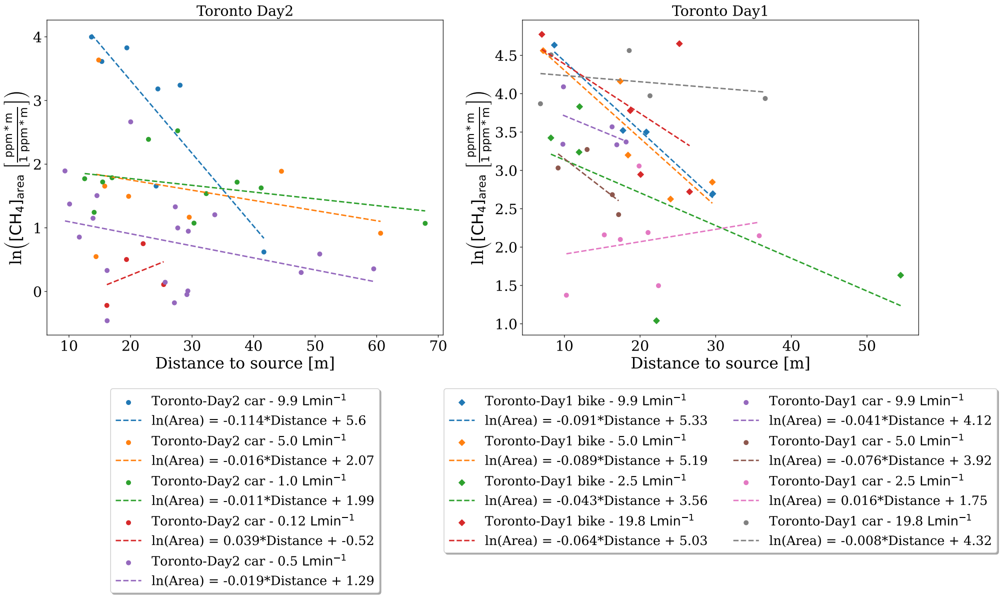


Figure S14. Logarithmic spatial peak area as function of distance to source for individual CH_4 enhancements reported in the Toronto Day 1 (a) and Day 2 (b) controlled release experiment. Colours represent different release rates, with a separate linear regression fitted for each rate.

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)):

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)).

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)). 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:

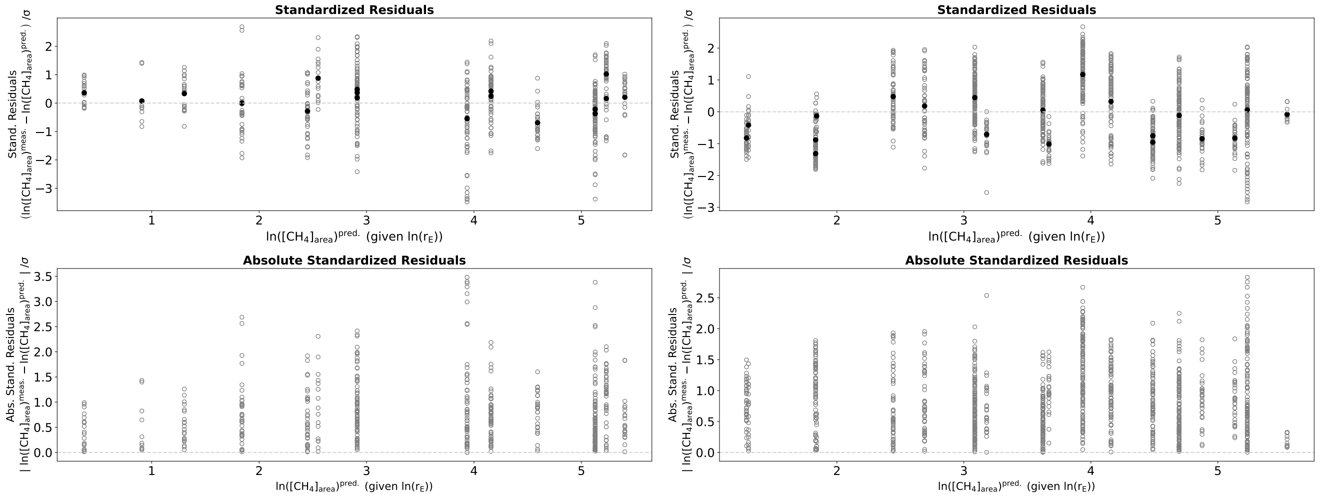
$$\frac{\ln([\mathrm{CH_4}]_{\mathrm{area}})^{\mathrm{measured}} - \ln([\mathrm{CH_4}]_{\mathrm{area}})^{\mathrm{predicted}}}{\sigma} \tag{6}$$

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 $\ln([\mathrm{CH_4}]_{\mathrm{area}})$) for the given values of the independent variable ($\ln(\mathrm{r_E})$). 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)). 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)). 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)). 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*).

Figure S15 illustrates the standardized residuals for the area linear regression model. The x-axis displays the predicted values of $\ln([\mathrm{CH_4}]_{\mathrm{area}})^{\mathrm{predicted}}$), i.e. the vertical point clouds represent the different release rates, but plotted here in terms of the corresponding $\ln([\mathrm{CH_4}]_{\mathrm{area}})$ 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 $\mathrm{Lmin^{-1}}$ (Figure S15b, $\ln([\mathrm{CH_4}]_{\mathrm{area}}) = 4.7$), 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 $\mathrm{Lmin^{-1}}$ in Utrecht I or 80 $\mathrm{Lmin^{-1}}$ in Rotterdam (Figure S15b, corresponding to an $\ln([\mathrm{CH_4}]_{\mathrm{area}})$ 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. Nonetheless, this trend is marked by significant fluctuations (Figure 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 Table S2 to Table S6. 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 (Table S2). 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 passed both tests (Table S3). 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 (Table S4). 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 (Table S5). 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 (Table S6).

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 S2. 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 S3. Utrecht: 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 S4. 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 S5. London: 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 S6. 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

Figure S16 to Figure 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 Lmin^{-1} releases appear to exhibit a bimodal shape (Figure S16a). This is also reflected in the QQ plots of the 20 and 80 Lmin^{-1} release rates (Figure 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).

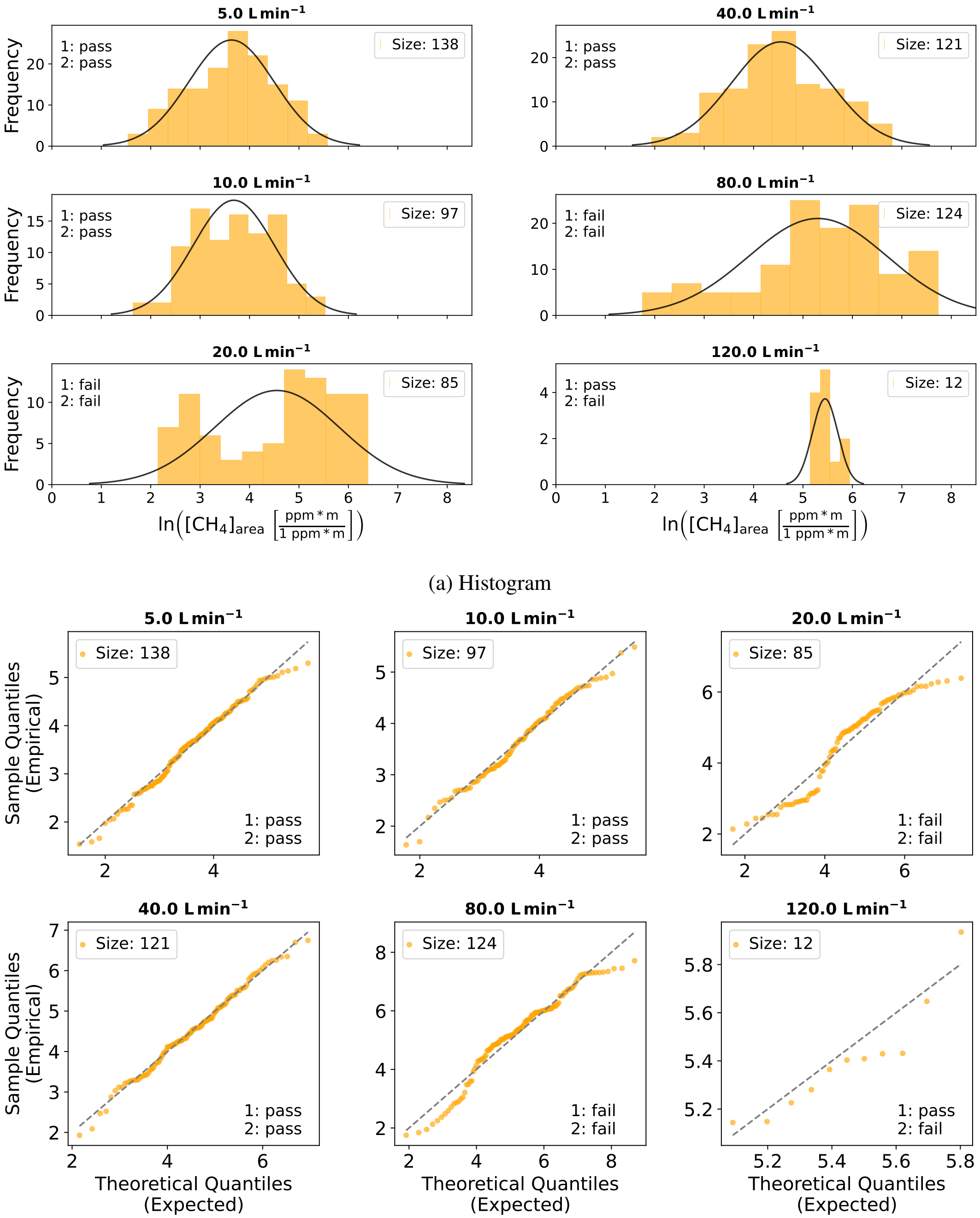
Both the histogram and QQ plot of the 2.18 Lmin^{-1} release in Utrecht confirm the positive assessments of both normality tests (Figure S17). However, the 3 Lmin^{-1} QQ plot exhibit an s-shaped form, confirming the fat tails visible in the histogram plot. The 15 Lmin^{-1} release rate distribution 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 (Figure S18).

The 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 (Figure 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 in the two test statistics, the visual observation of the histogram and QQ plots also shows normality in almost all cases (Figure 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.

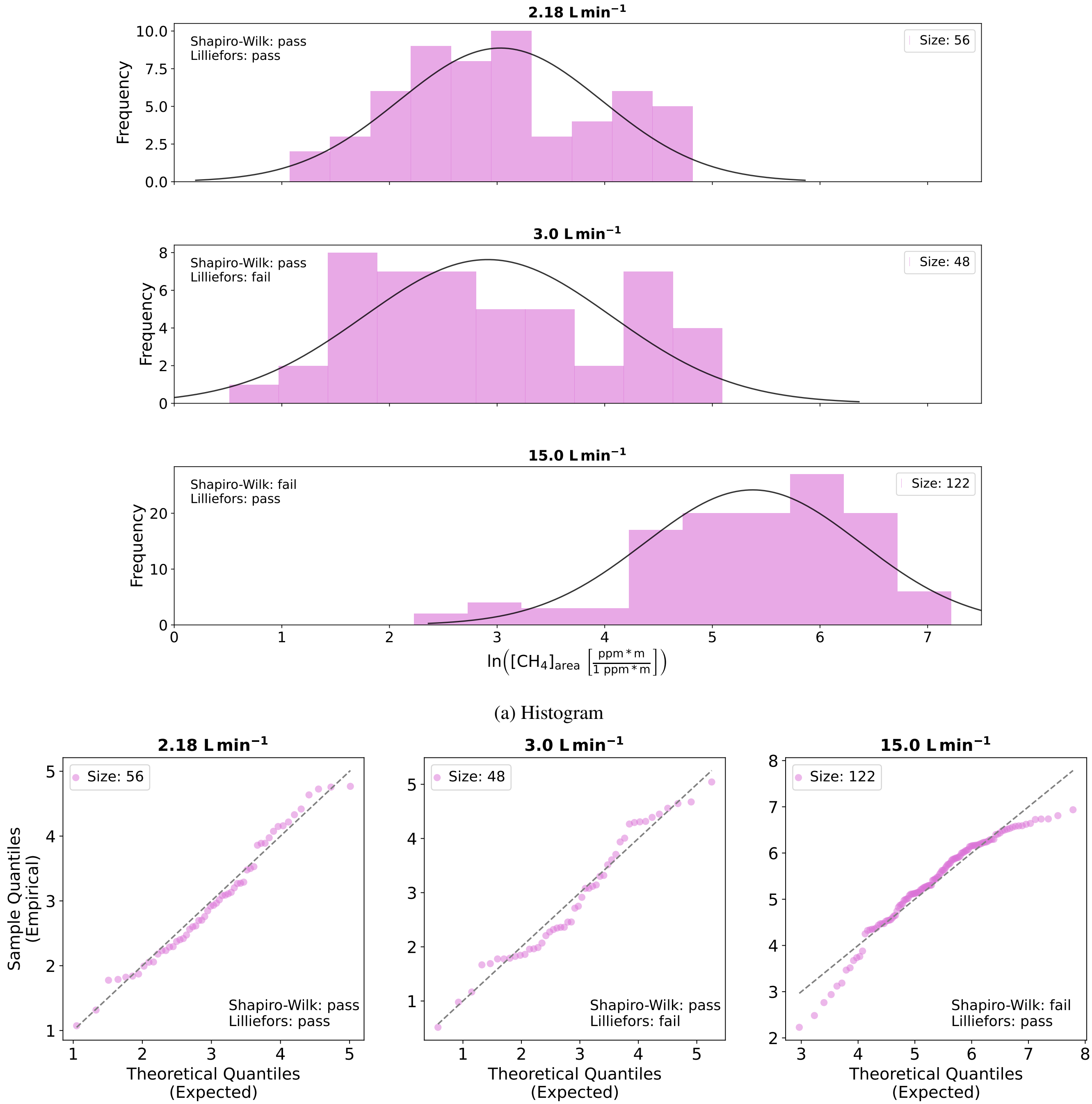
Generally, as the release rates increase, a shift of the centre of the distributions towards higher $\ln([\mathrm{CH}_4]_{\mathrm{area}})$ values is observed. For the majority of $\ln([\mathrm{CH}_4]_{\mathrm{area}})$ distributions the QQ plots suggest normality, confirming the evaluation of the test 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

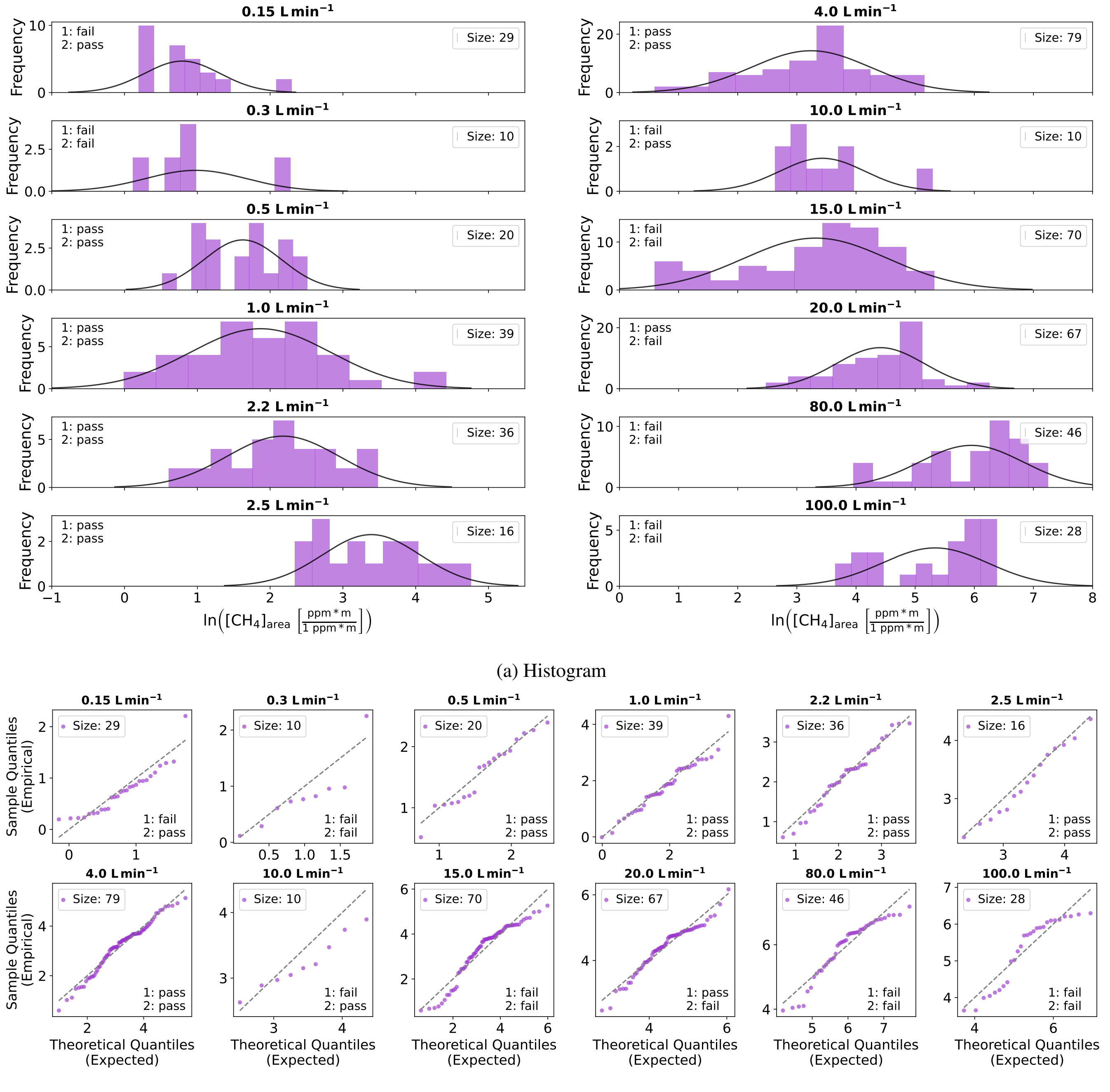
Figure S16. Rotterdam: Assessment of log-normality. (a) Histogram of the logarithmically transformed integrated peak area of the measured CH_4 enhancements. Each histogram represents areas measured at a given release rate. A Gaussian distribution is fitted to the data and the results of two normality 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 ($\ln([\mathrm{CH}_4]_{\mathrm{area}})$) versus a normal distribution for each release rate separately.



(a) Histogram

(b) Quantile-Quantile plot

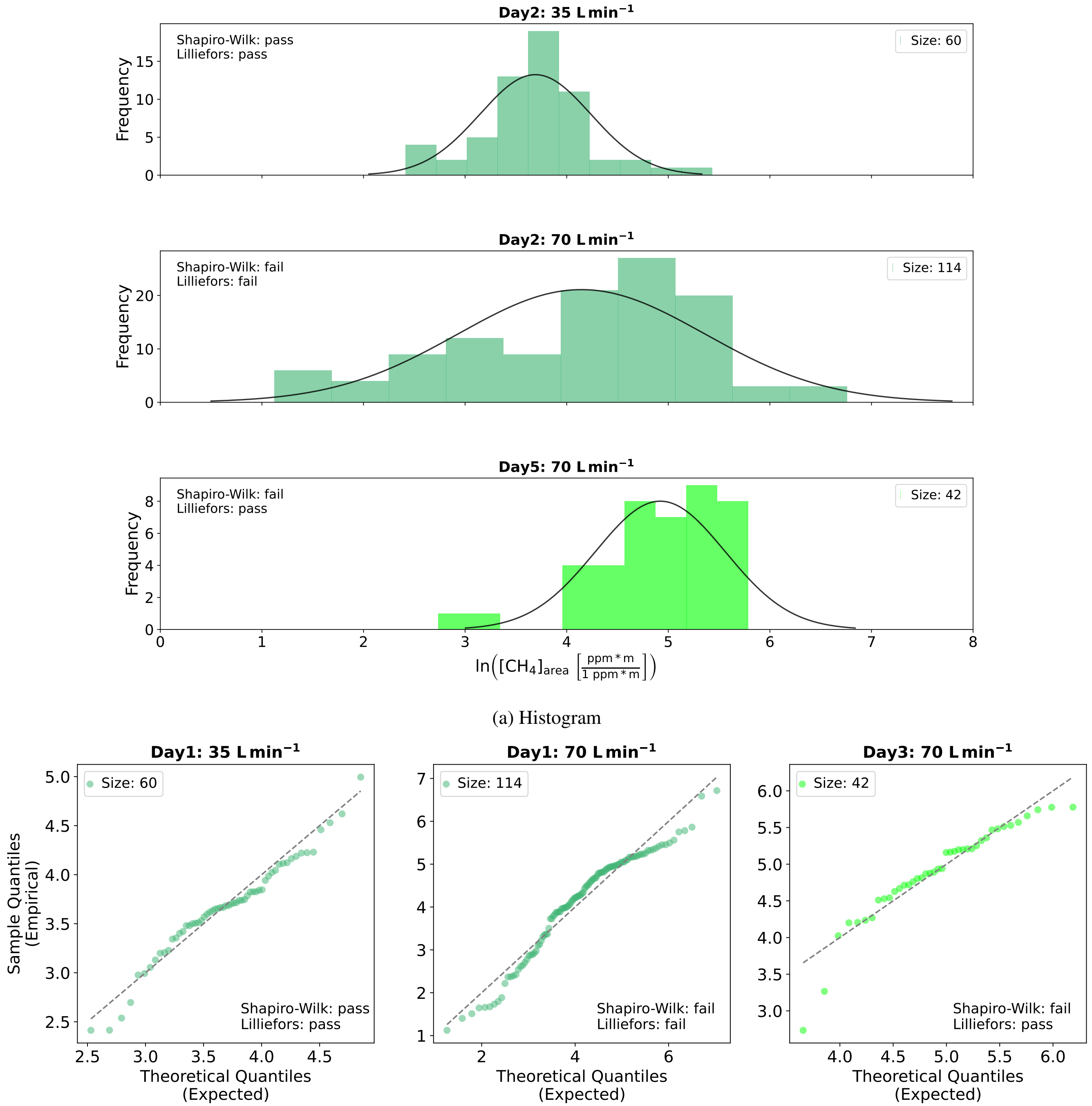
Figure S17. Utrecht: Assessment of log-normality. (a) Histogram of the logarithmically transformed integrated peak area of the measured CH_4 enhancements. Each histogram represents areas measured at a given release rate. A Gaussian distribution is fitted to the data and the results of two normality tests (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 ($\ln([\mathrm{CH}_4]_{\mathrm{area}})$) versus a normal distribution for each release rate separately.



(a) Histogram

(b) Quantile-Quantile plot

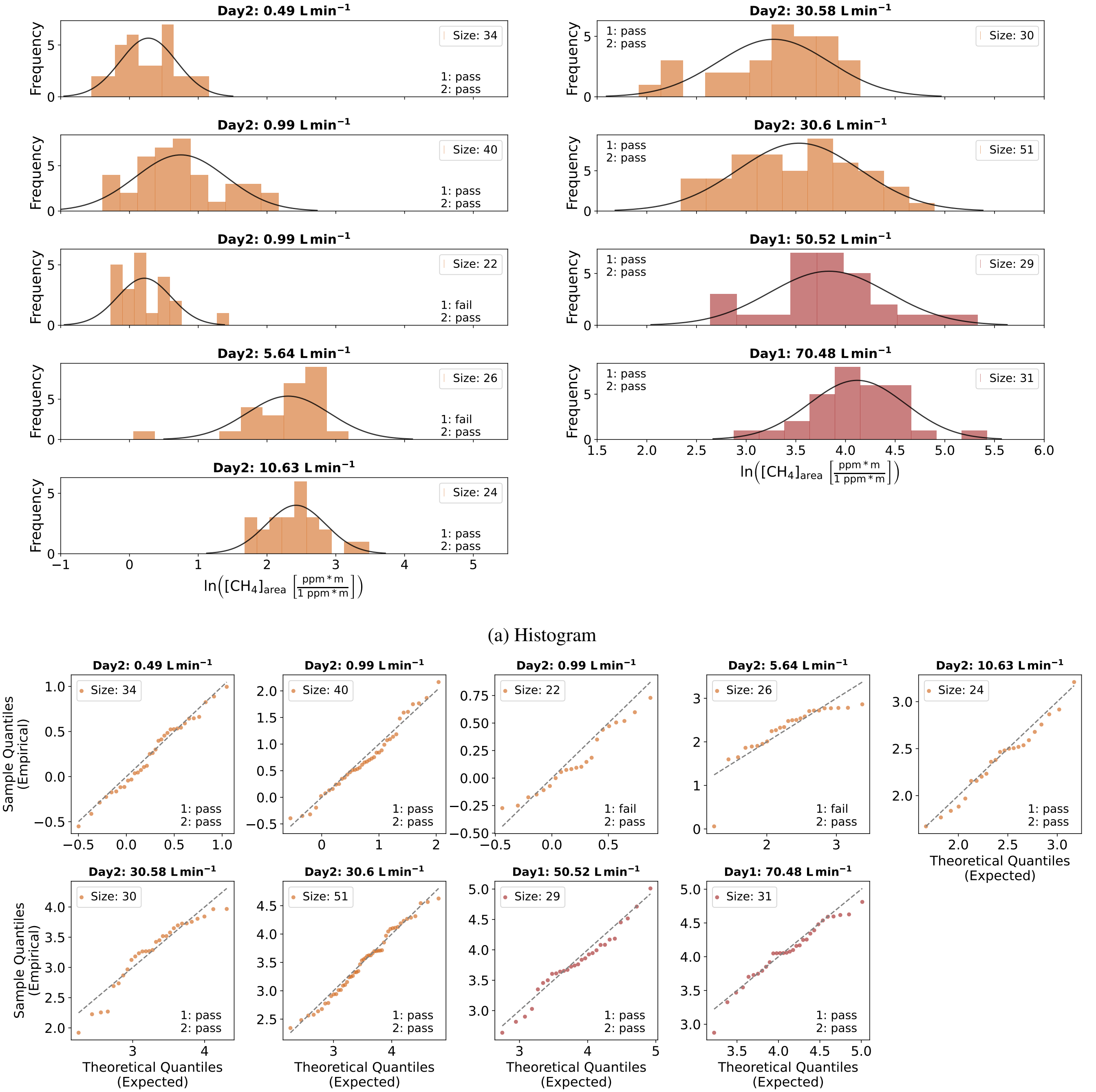
Figure S18. Utrecht II: Assessment of log-normality. (a) Histogram of the logarithmically transformed integrated peak area of the measured CH_4 enhancements. Each histogram represents areas measured at a given release rate. A Gaussian distribution is fitted to the data and the results of two normality 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: Assessment of log-normality. (a) Histogram of the logarithmically transformed integrated peak area of the measured $\mathrm{CH_4}$ 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 ($\ln([\mathrm{CH_4}]_{\mathrm{area}})$) versus a normal distribution for each release rate separately.



(a) Histogram

(b) Quantile-Quantile plot

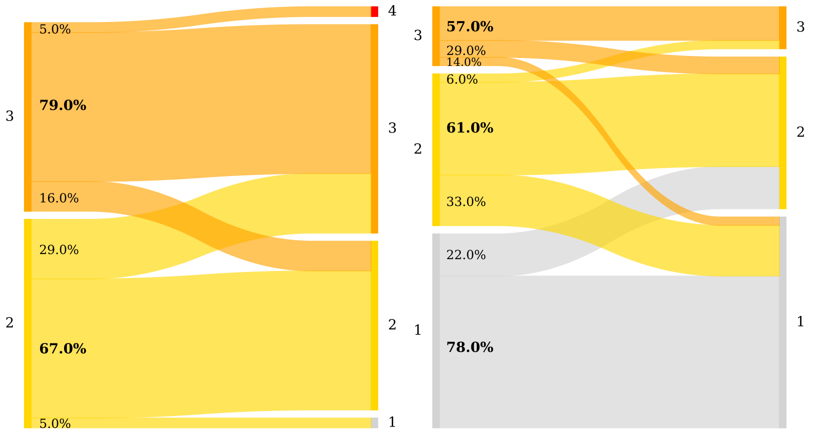
Figure S20. London II: Assessment of log-normality. (a) Histogram of the logarithmically transformed integrated peak area of the measured CH_4 enhancements. Each histogram represents areas measured at a given release rate. A Gaussian distribution is fitted to the data and the results of two normality 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 ($\ln([\mathrm{CH}_4]_{\mathrm{area}})$) versus a normal distribution for each release rate separately.

S8 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 Table S7 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 Lmin^{-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 Figure S22 and Figure S21.

Table S7. Natural gas distribution network CH_4 emission categories. Corresponding maxima ranges and area ranges are given for the two emission rate estimation methods.

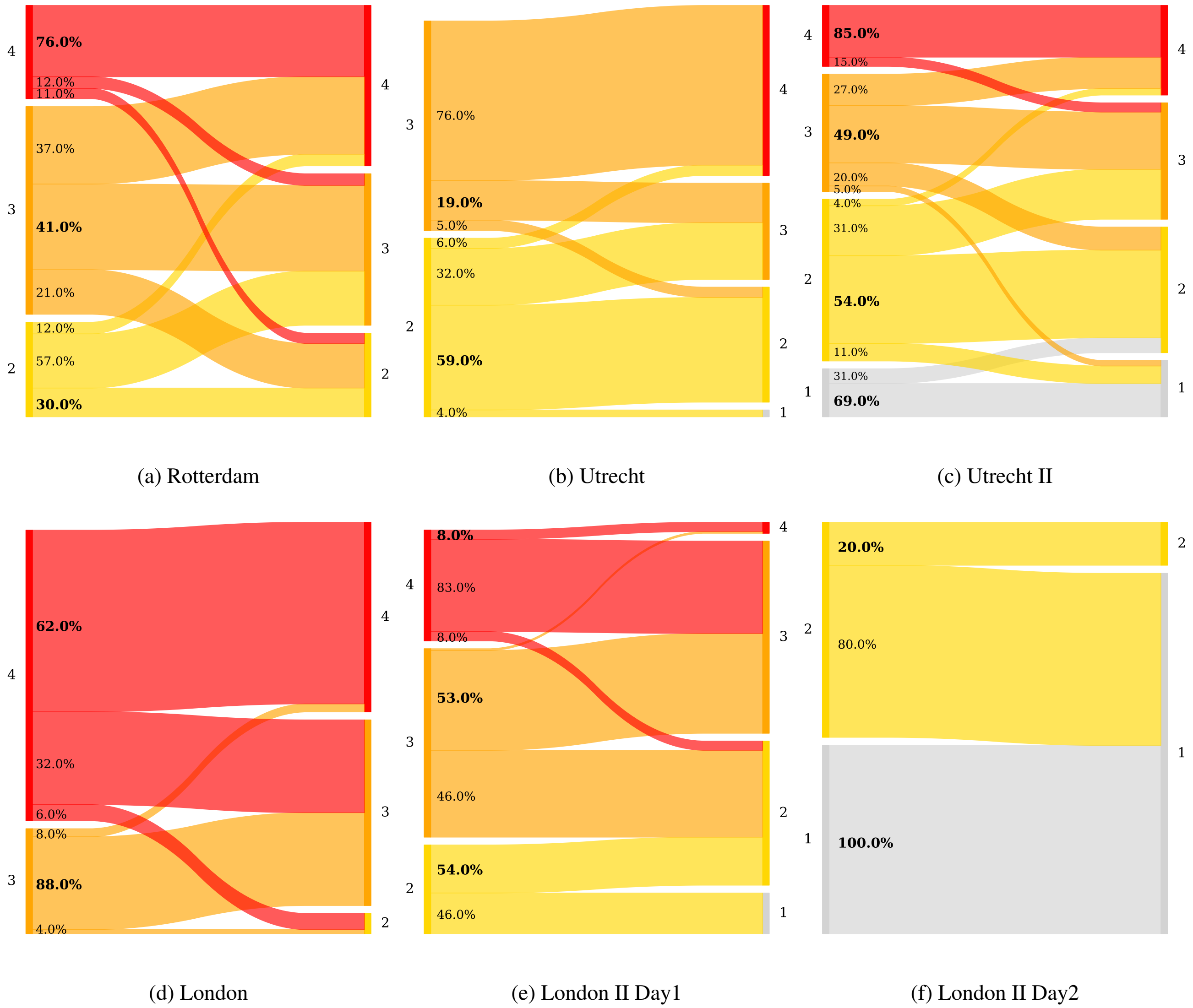
Class	Emission Rate $\left[\mathrm{Lmin}^{-1}\right]$	Weller eq. $[\mathrm{ppm}]$	Area eq. $[\mathrm{ppm} * \mathrm{m}]$
High	> 40	> 7.6	> 109
Medium	$6 - 40$	$1.6 - 7.6$	$25 - 109$
Low	$0.5 - 6$	$0.2 - 1.59$	$3.7 - 25$
Very Low	< 0.5	< 0.2	< 3.7



(a) Toronto Day1

(b) Toronto Day2

Figure S21. 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.



(a) Rotterdam

(b) Utrecht

(c) Utrecht II

(d) London

(e) London II Day1

(f) London II Day2

Figure S22. Categorization performance for data obtained in Rotterdam, Utrecht, Utrecht II, London 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.

S9 Influence of Sampling Effort

S9.1 Hypothetical Distributions

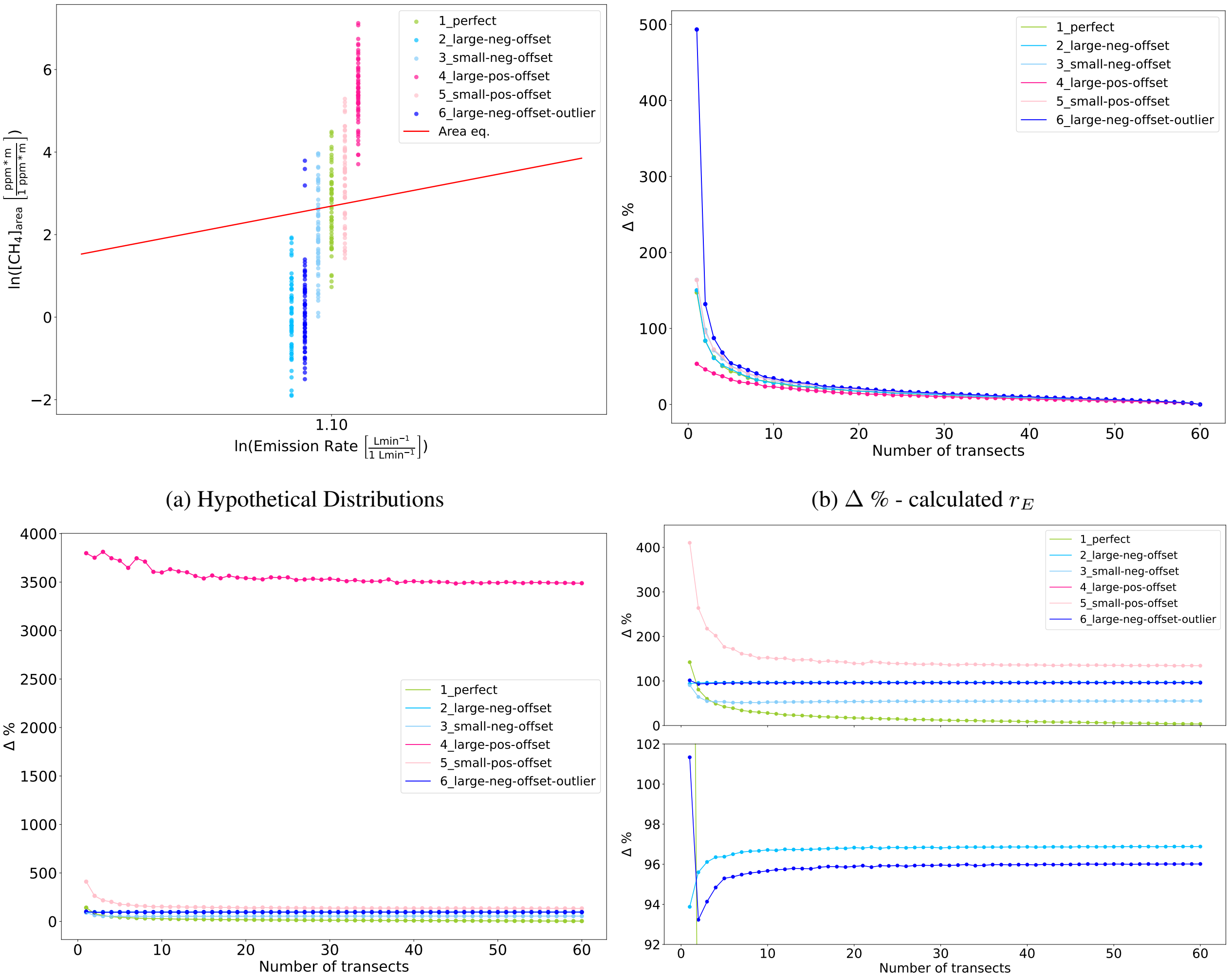
In order to illustrate the behaviour of sampling multiple times at the same locations, we present results for some selected hypothetical distributions with means falling above or below the empirical equation derived from the totality of all measurements in the main manuscript. Figure S23 and Figure S24 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 Figure S23b decreases to 0 as expected. 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 Figure S24b decreases to 0 as expected, except for the distribution with the high positive offset. 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.

Figure S25 and Figure S26 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, 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.



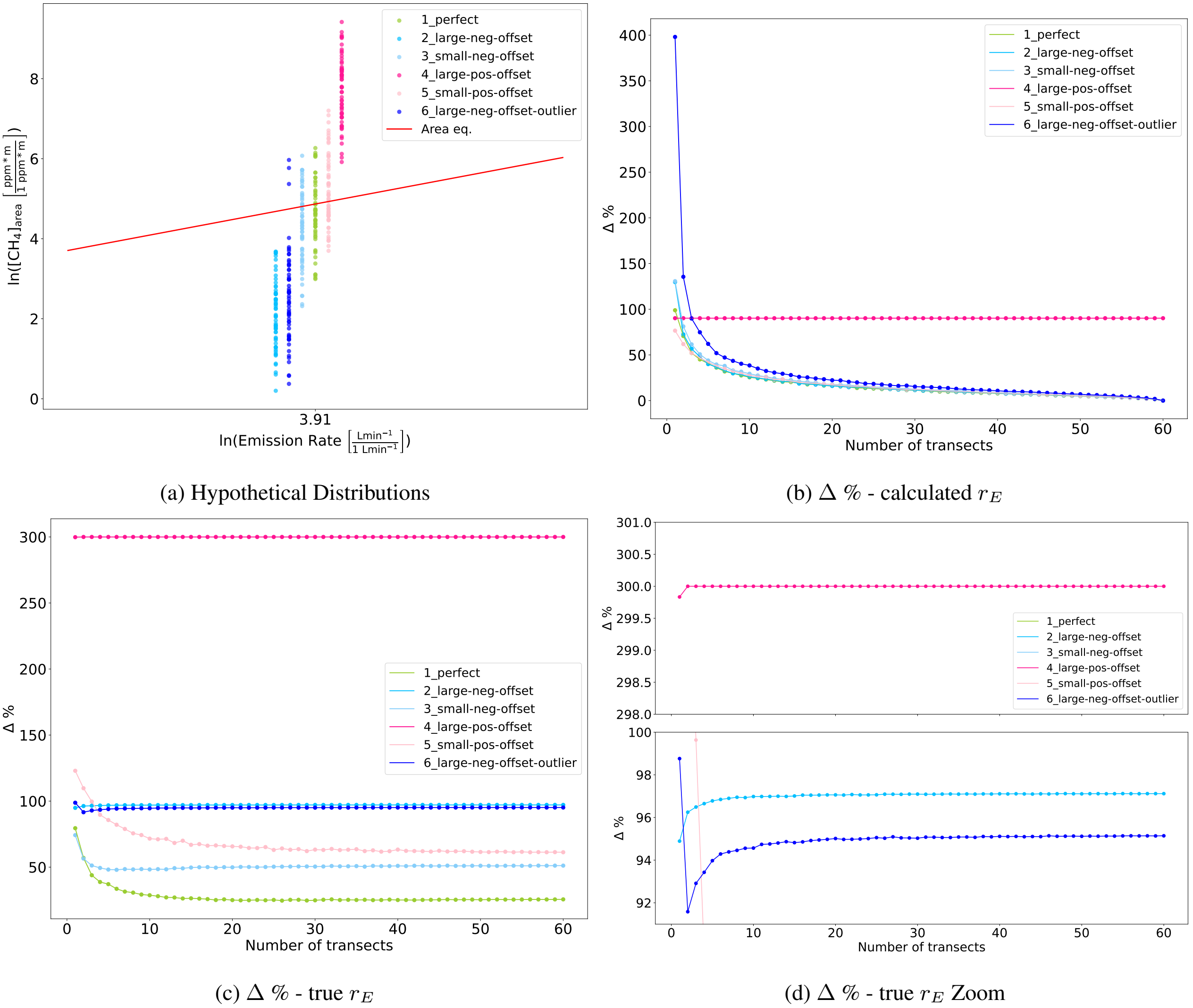
(a) Hypothetical Distributions

(b) Δ % - calculated r_E

(c) Δ % - true r_E

(d) Δ % - true r_E Zoom

Figure S23. Hypothetical distributions for a release rate 3 Lmin^{-1} with different offsets from the perfect distribution.



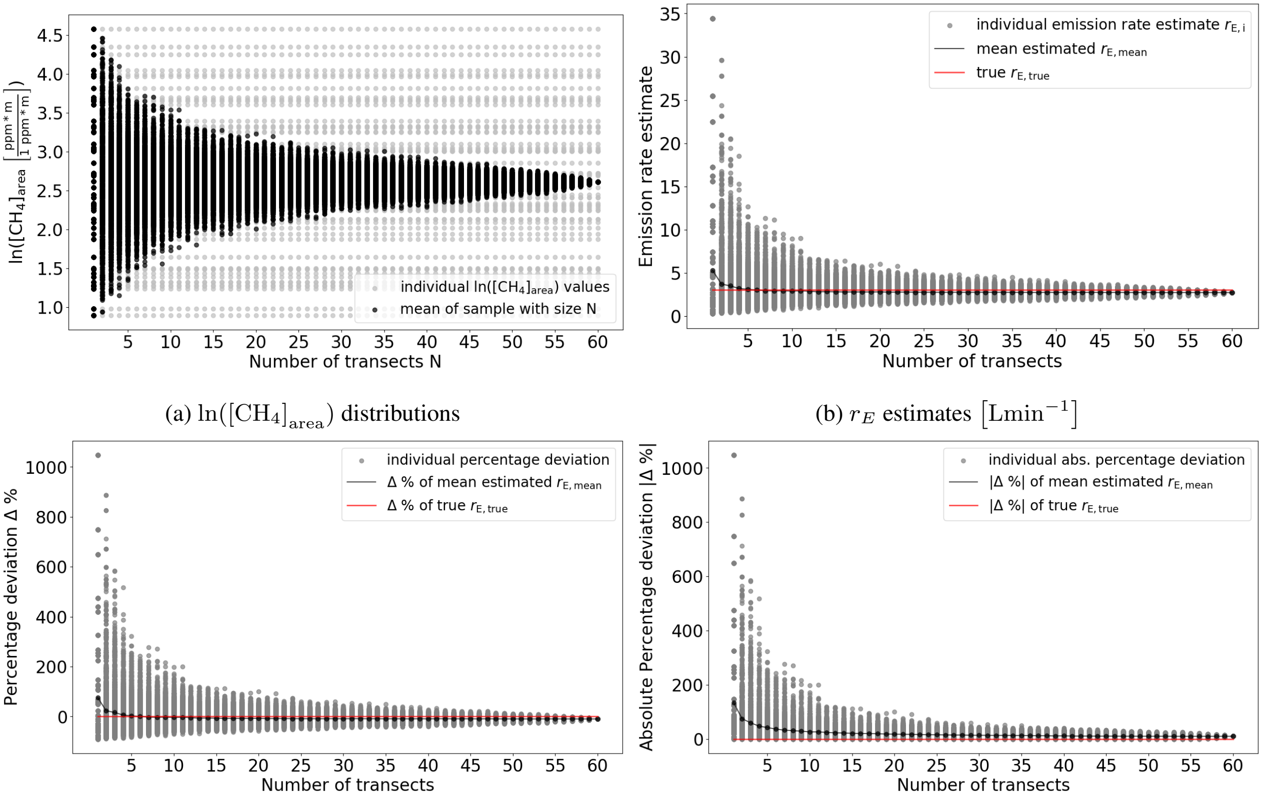
(a) Hypothetical Distributions

(b) Δ % - calculated r_E

(c) Δ % - true r_E

(d) Δ % - true r_E Zoom

Figure S24. Hypothetical distributions for a release rate 50 Lmin^{-1} with different offsets from the perfect distribution.



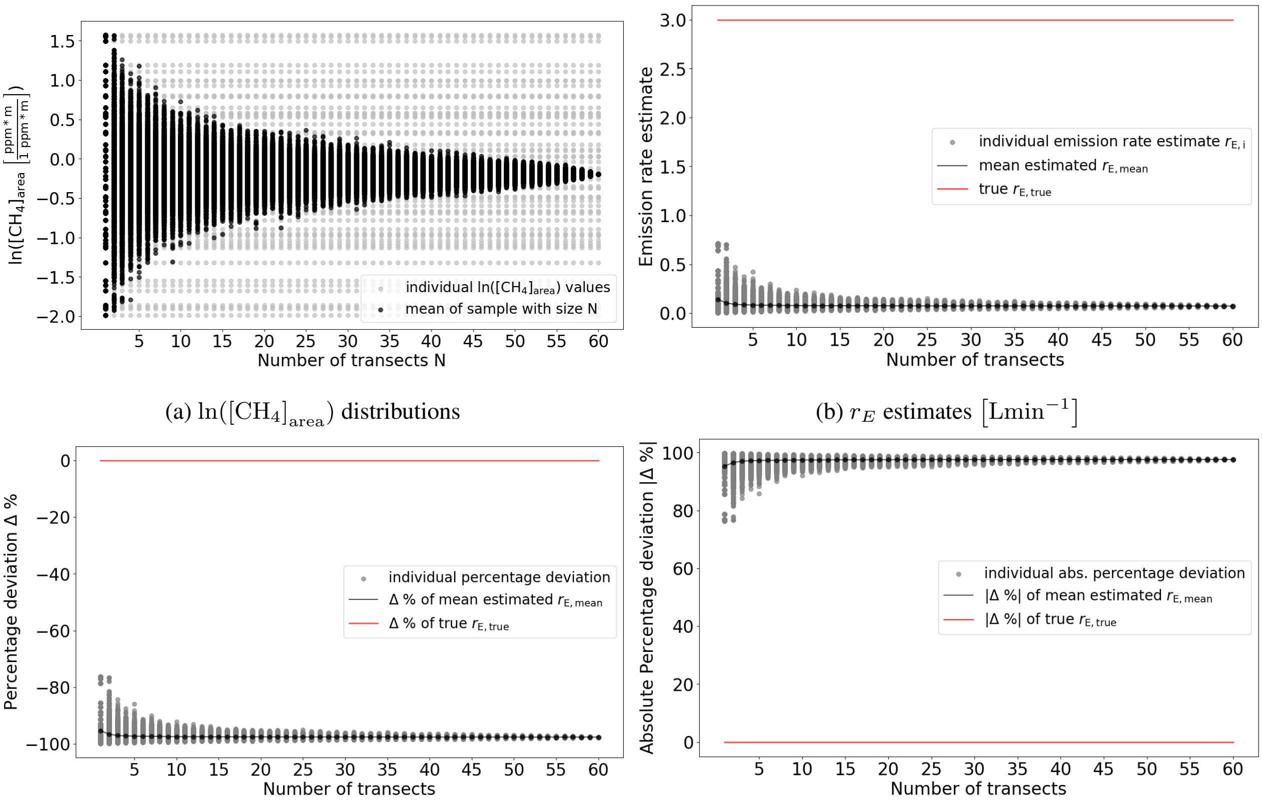
(a) $\ln([\mathrm{CH}_4]_{\mathrm{area}})$ distributions

(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 S25. 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 $[\mathrm{Lmin^{-1}}]$

(c) Δ % - r_E from true r_E

(d) Absolute Δ % - r_E from true r_E

Figure S26. 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: 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, 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, 2006.

Nimon, K.: Statistical Assumptions of Substantive Analyses Across the General Linear Model: A Mini-Review, Frontiers in Psychology, 3, 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, 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, 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.