Source Attribution and Emissions Quantification for Methane Leak Detection: A Non-Linear Bayesian Regression Approach

Mirco Milletarì[1], Sara Malvar[1], Yagna D. Oruganti[1], Leonardo O. Nunes[1], Yazeed Alaudah[1], and Anirudh Badam[1]

Microsoft

Abstract. Methane leak detection and remediation efforts are critical for combating climate change due to methane's role as a potent greenhouse gas. In this work, we consider the problem of source attribution and leak quantification: given a set of methane ground sensor readings, our goal is to determine the sources of the leaks and quantify their size in order to enable prompt remediation efforts and to assess the environmental impact of such emissions. Previous works considering a Bayesian inversion framework have focused on the over-determined (more sensors than sources) regime and a linear dependence of methane concentration on the leak rates. In this paper, we focus on the opposite, industry-relevant regime of few sources per sensor (under-determined regime) and consider a non-linear dependence on the leak rates. We find the model to be robust in determining the location of the major emission sources, and their leak rate quantification, especially when the signal strength from the source at a sensor location is high.

Keywords: Bayesian framework · source attribution · inverse problem · leak quantification.

1 Introduction

Methane (CH_4), the primary component of natural gas, is a potent greenhouse gas (GHG) with a Global Warming Potential (GWP) of 84–87 over a 20-year timescale [6]. The Intergovernmental Panel on Climate Change (IPCC) affirms that reduction of anthropogenic methane emissions is the most efficient way to curb a global temperature rise of 1.5°C above pre-industrial levels by 2030 [17].

The global oil and gas industry is one of the primary sources of anthropogenic methane emissions, with significant leaks occurring across the entire oil and gas value chain, from production and processing to transmission, storage, and distribution. Examples of sources of methane leaks are malfunctioning clamps, flares, flow lines, tanks, pressure regulators, thief hatches, and valves. Capacity limitations in gathering, processing, and transportation infrastructure can also lead to the venting of excess methane. The International Energy Agency (IEA) estimates [10] that it is technically possible to avoid around 70% of today's methane emissions from global oil and gas operations. These statistics highlight the importance of leveraging various methane detection technologies and source attribution techniques to address this critical issue.

Most of these technologies rely on complex models of particulate transport in the atmosphere. Complexity is due to the interplay of multiple spatial scales (from the particle scale to near-source and long-range effects), multi-physics (coupling mass transport, turbulence, chemistry, and wet/dry deposition), and complex geometry (e.g., flow over topography or man-made structures). Atmospheric dispersion models have a long history, reaching back to Richardson's [16] and Taylor's [21] pioneering investigations of turbulent diffusion. However, maintaining accuracy is a prevalent challenge in dispersion modeling since many models have large uncertainties in effective parameters, such as the Monin-Obukhov length [15], atmospheric stability classes, or terrain roughness length.

Past research has mostly focused on improving forward transport models to evaluate downstream pollutant concentrations given source leak rates and meteorological variables. However, few works have focused on the source attribution problem, which belongs to the class of inverse problems. Methods for estimating source strength and/or location from measurements of concentration can be divided into two major categories depending on the physical scale of the problem. Researchers employed ground-based measurements and a high-resolution mesoscale air transport model to quantify GHG emissions at the urban, regional, and continental scales. They use a Bayesian statistical technique to predict emissions and the associated uncertainty by combining previous emission inventories with atmospheric measurements [14]. When the physical distance between the sources and sensor observations is minimal, using mesoscale air transport models for inversion becomes challenging. Typically, at such scales, atmospheric inversions are performed using plume dispersion and surface layer models. In this paper, we are primarily interested in observations taken relatively close to the source, and we limit ourselves to analyzing the uncertainty in inverse modeling linked to plume dispersion models.

A considerable number of inversion studies based on plume inversion models have been published in peer-reviewed journals. Several of these papers deal with uncertainty estimations [11, 18]. Garcia et al. [8], for instance, considered a Bayesian regression model using a non-stationary forward operator while Lushi and Stockie [13] considered a positively constrained, linear least squared method together with the Gaussian Plume model to determine the leak rates of the sources. However, both studies considered a linear dependence on the leak rates and a design where the number of sensors (9) is much greater than the number of sources (4). Mathematically, the latter scenario results in an over-determined system, for which Linear Programming solvers work well.

In this paper, we propose a solution based on Bayesian optimization to identify the source of a methane leak and to quantify the size of the leak, using readings from a spatially sparse array of sensors, which corresponds to a mathematically under-determined system. This is a particularly relevant scenario for many industrial applications that require monitoring of large areas with costly ground sensors. The scenario that we consider in this study is for continuous monitoring of an Area of Interest (AoI), where an operator would be interested in detecting anomalous methane leaks, identifying their likely sources, and estimating leak size in near-real-time, to allow for prompt inspection and remediation. As a result, the proposed methodology focuses on achieving a reasonable trade-off between accuracy and a relatively low computational time.

2 Methods

Source attribution belongs to the class of inverse problems; it aims at finding the sources that generated a certain field configuration given readings of field values at some restricted number of points $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_M\} \in \mathbb{R}^d$, where d is the number of space dimensions. In this work, we consider the following scenario: during some observation time δt, some or all the sensors deployed in the field record methane concentration signals, exceeding a determined threshold. As there are multiple sources being monitored in the field, the sensors only record a compound signal that is assumed to be given by the linear combination of concentrations generated by a subset of sources at the each sensor location. The objective is therefore to find the decomposition of the compound signal to determine the contribution of each source. We are interested in determining the k sources that contribute the most to the signal and estimate their strength.

2.1 Bayesian approach

The Bayesian approach relies on inverting the forward model using Bayes' principle and sampling algorithms, based on some form of Markov Chain Monte Carlo (MCMC) or Stochastic Variational Inference (SVI). In a physical model, all empirical parameters are subject to systematic and statistical errors; the former considers the measurement error associated with the instrument(s), while the latter encompasses statistical uncertainty in a set of measurements. In a Bayesian approach, this input uncertainty naturally propagates through the model in a non-parametric fashion. As a result, inferred parameters come with confidence levels that better reflect the physical reality of the model. This means that all quantities are expressed by probability distributions rather than single numbers. In general, given a parameter set $\boldsymbol{\theta}$, a variable set $\boldsymbol{q}$, and sensor readings $\boldsymbol{w}$, Bayes' principle reads: $P(\boldsymbol{q}, \boldsymbol{\theta}|\boldsymbol{w}) = P(\boldsymbol{w}|\boldsymbol{q}, \boldsymbol{\theta})P(\boldsymbol{q}, \boldsymbol{\theta})/Z(\boldsymbol{w})$, where $P(\boldsymbol{q}, \boldsymbol{\theta})$ is the prior, based on our current knowledge or assumptions on the form of the distribution, $P(\boldsymbol{w}|\boldsymbol{q}, \boldsymbol{\theta})$ is the likelihood, and $P(\boldsymbol{q}, \boldsymbol{\theta}|\boldsymbol{w})$ the posterior. Finally, $Z(\boldsymbol{w})$ is a normalization. In this work we restrict our analysis to a scenario where all the source locations are known. In this case, the unknowns are the leak rates of the N sources $\boldsymbol{q} = [q_1, q_2, \cdots, q_N]^T$ measured in kg/h. The methods presented here can be extended to scenarios with known and unknown sources, where the latter was considered in Wade and Senocak [24].

Let us call $A_{mn}(\mathbf{q}, \boldsymbol{\theta})$ the $M \times N$ (M sensors and N sources, $M \ll N$) forward operator mapping the concentration field from source to sensor location, parametrized by $\boldsymbol{\theta} = \{u, \phi, p\}$, where u is the modulus of the wind velocity $[m\, s^{-1}]$; ϕ its in-plane direction $[rad]$; and possibly other p parameters that depend on the details of the model. While $\boldsymbol{\theta}$ are measured quantities (with uncertainties), $\boldsymbol{q}$ are unknown and constitutes the fitting parameters of the model. Given an array of M sensors, let us call W_m the compound signal recorded at sensor m at time intervals δt. Then we have the relation:

$$w_m = \sum_{n=1}^{N} A_{mn}(q_n, \boldsymbol{\theta}) \equiv \mathcal{A}(\boldsymbol{q}, \boldsymbol{\theta}). \tag{1}$$

In general, this is a non-linear, time-dependent mapping, solution of the Diffusion-Advection partial differential equation (PDE). Following [8], we write it as:

$$\big(\partial_t + \boldsymbol{L}(\theta)\big) C(\boldsymbol{x}, t) = \sum_{n=1}^{N} q_n(t) \delta(\boldsymbol{x} - \boldsymbol{x}_n) \tag{2}$$

$$\boldsymbol{L}(\theta) = \boldsymbol{\nabla} \cdot \big(\boldsymbol{u}(\boldsymbol{x}, t) - \boldsymbol{D}(\boldsymbol{x}) \boldsymbol{\nabla}\big), \tag{3}$$

where $\boldsymbol{L}(\boldsymbol{\theta})$ is a linear operator, possibly depending non-linearly on the parameters $\boldsymbol{\theta}$, comprising an advection and a diffusion term controlled by the diffusion matrix, $\boldsymbol{D}(\boldsymbol{x})$. The term $C(\boldsymbol{x}, t)$ is the concentration field at location $\boldsymbol{x} = (x, y, z)$ and time t. Finally, note that we are considering point-emission sources specified by the $\boldsymbol{x}_n$ coordinates in the Dirac delta function on the right hand side of Eq. (2). The solution implemented in the next section imposes a series of assumptions on the form of $C(\boldsymbol{x}, t)$ and, therefore, of $\mathcal{A}$ that makes the problem numerically manageable at different levels of complexity.

2.2 Non-linear Bayesian regression: stationary model

To simulate the contribution of each source, we consider a forward operator based on the Gaussian plume model [23], which is a special solution of Eqs. (2) and (3) under the following simplifying assumptions:

1. The leak rates, $\boldsymbol{q}(t)$, vary slowly in time such that it can be considered constant over the measurement time scale, i.e. $\boldsymbol{q}(t) = \boldsymbol{q}$.
2. The wind velocity and direction are stationary and aligned along the x direction for $x \geq 0$, i.e. $\boldsymbol{u} = (u, 0, 0)$.
3. The diffusion matrix, $\boldsymbol{D}(\boldsymbol{x})$, is replaced by effective parameters based on the Pasquill stability class.

Boundary conditions include finiteness of the concentration field at the origin and infinity, together with the condition that the contaminant does not penetrate the ground, see [20] for details. Under these conditions, the PDE admits an analytical solution in the form of a Gaussian kernel:

$$C_n(\boldsymbol{x}) = \frac{q_n}{2\pi\, u\, \sigma_y\, \sigma_z} \exp\left\{ -\frac{(z-h)^2}{2\sigma_z^2} - \frac{(z+h)^2}{2\sigma_z^2} - \frac{y^2}{2\sigma_y^2} \right\}, \tag{4}$$

with the (scalar) concentration field measured in [kg m^{-3}], although we will often convert this to parts per million per volume (ppmv) in the rest of the paper. The σ_i are standard deviations, and h is the height of the source. Our implementation of the Gaussian plume model follows the one implemented in the `Chama`[1] open-source library [12], where the value of the standard deviations is re-defined to include heuristic information on the stability of the plume: $\sigma_i(x) = a_i\, x\,(1 + x\, b_i^{-1})^{-c_i}$, where the values of the parameters a_i, b_i, c_i depend on the atmospheric stability class (indexed from A to F) and are different for the y and z components. Weather stability classes are evaluated given the surface wind, cloud coverage, and the amount of solar radiation in the AoI on a specific date and time. Wind direction and source location are re-introduced respectively by rotating the simulation grid in-plane and by re-centering it on the source position. This expression, linear in the leak rate $\boldsymbol{q}$, was used in [13] as the diffusion/advection operator of a linear regression model. Following `Chama` [12], we consider buoyancy corrections to dispersion along the z-axis, introduced heuristically in Eq. (4) as:

$$z'_n = z + 1.6\,\frac{B_n^{1/3}\, x^{2/3}}{u}, \quad B_n = \frac{g\, q_n}{\pi}\left(\frac{1}{\rho_{CH_4}} - \frac{1}{\rho_{air}}\right). \tag{5}$$

Where g is the gravitational constant while ρ_{CH_4} and ρ_{air} are the density of methane and air measured in standard conditions. As such, we measure buoyancy in units of $m^4\, s^{-3}$. Note that this transformation introduces a non-linear dependence on $\boldsymbol{q}$, making our source attribution model non-linear.

To accelerate the Gaussian plume model for large-scale simulations, we leverage `PyTorch`[2] to parallelize evaluation over both sensors and sources. This allows for a massive speedup of over 50 times (on CPU) compared to the `Chama` implementation, with further speed-up possible by leveraging GPUs. Moreover, this enables gradient evaluation of training parameters in the model (in our case the leak rates and possibly the atmospheric data) necessary for the Bayesian optimization process using the Hamiltonian Montecarlo algorithm provided by the open-source Bayesian optimization library, `Pyro` [3] [4].

Assuming a normal distribution of the noise with covariance matrix Σ, the likelihood of the model reads:

$$P(\boldsymbol{w}|\boldsymbol{q},\boldsymbol{\theta}) = \frac{1}{(2\pi\, \det \Sigma)^{1/2}} e^{-\frac{1}{2}||\Sigma^{-1/2}(\boldsymbol{w}-\mathcal{A}(\boldsymbol{q},\boldsymbol{\theta}))||^2}. \tag{6}$$

Due to the $M \ll N$ regime we are interested in, corresponding to a low-density sensor placement, the problem is under-determined. While the parameters $\boldsymbol{\theta}$ depends on atmospheric conditions, the leak rates depend on the specifics of the physical process that led to the emission. Therefore, following García et al. [8], we reasonably assume $\boldsymbol{\theta}$ and $\boldsymbol{q}$ to be statistically independent; as a consequence, the prior distribution factorizes as $P(\boldsymbol{q},\boldsymbol{\theta}) = P(\boldsymbol{q})P(\boldsymbol{\theta})$, where the distribution on $\boldsymbol{\theta}$

[1] `https://github.com/sandialabs/chama`

[2] `https://github.com/pytorch/pytorch`

[3] `https://github.com/pyro-ppl/pyro`

is obtained via direct measurement of the weather data at a particular location, together with their experimental (and possibly statistical) uncertainties due to temporal or spatial averaging. However, in the application considered in the next sections, we will make the simplifying assumption of $\boldsymbol{\theta}$ being deterministic; the reason for this choice is related to the considered industrial scenario, see Sect. 3 for details. In comparison, the scenario considered in Garcia [8] and Lushi [13] dealt with the detection of lead-zinc emission; in this case, sensors need to collect enough samples from direct deposition of the pollutants in a collection device, which may require hours, depending on the deposition velocity of the pollutant. In this case, ten minute averages of wind data were used over the measurement time. In the case of methane, sensors operate in near real-time using direct measurements based on a variety of techniques, such as mid/near infra-red lasers or metal oxide semiconductors to name a few. As we detail in Sect. 3 below, we are interested in a near real-time source attribution scenario; in this case, weather data will be taken at the time of measurement from the sensor's weather stations. By taking $\boldsymbol{\theta}$ as deterministic, we are therefore assuming that weather data are homogeneous across the AoI, in agreement with the same assumption used to obtain the Gaussian plume solution, and neglect systematic errors.

2.3 Ranking Model

To rank the source contributions, we use multi-point estimates of each leak rate's posterior distribution. As estimators, we take percentiles from 0 to 100 at steps of 2 lying in the 68% HPDI confidence interval, plus the sample average. Sampled point estimates are then used to reconstruct the source contribution to the signal measured at each sensor using again the forward model. For each prediction, we evaluate again the error with the observed value at each sensor location and evaluate the posterior predictive likelihood $\mathcal{P}_s \equiv P_s(\boldsymbol{w}|\boldsymbol{q}^\star_s, \boldsymbol{\theta})$, $q^\star_{sn}$ being the s-th point estimates of the n-th leak rate from the marginal posterior distribution; this will be used in the final ranking step to weight the goodness of the ranking solution. Note that we denote with $\star$ a variable or parameter fixed by a particular operation, e.g. optimization, sorting, or max.

Each one of the s samples from point estimates (also referred to as point samples) propose a different source reconstruction within the 68% HPDI of the marginal posteriors. By ranking emission sources by their contribution at each sensor, we obtain an ensemble of possible ranking:

$$R^s_{mn^\star} = \arg \underset{n}{\mathrm{sort}} \; A^s_{mn}(q^\star_{sn}, \boldsymbol{\theta}), \tag{7}$$

where s is the point sample index and $A^s_{mn}(q^\star_{sn}, \boldsymbol{\theta})$ is the methane concentration value of source n measured from sensor m, obtained from the point estimate s of the leak rate. Each member of the ranking ensemble is weighted by the related predictive likelihood. The final ranking is obtained as a composite estimator. For each sensor, we take the proposed ranking with the highest likelihood:

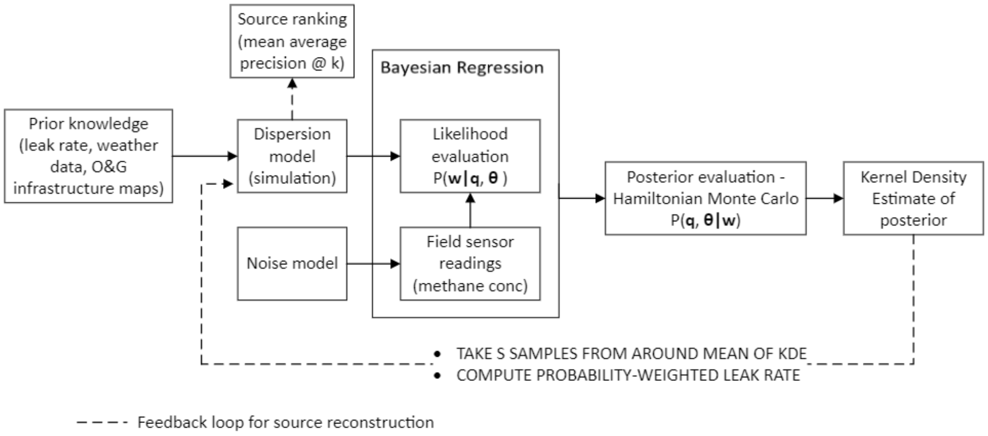

Source ranking (mean average precision @ k)
Bayesian Regression
Prior knowledge (leak rate, weather data, O&G infrastructure maps)
Dispersion model (simulation)
Likelihood evaluation P(**w**|**q**, **θ**)
Posterior evaluation - Hamiltonian Monte Carlo P(**q**, **θ**|**w**)
Kernel Density Estimate of posterior
Noise model
Field sensor readings (methane conc)
TAKE S SAMPLES FROM AROUND MEAN OF KDE
COMPUTE PROBABILITY-WEIGHTED LEAK RATE
Feedback loop for source reconstruction

Fig. 1: Flow chart of the source attribution methodology

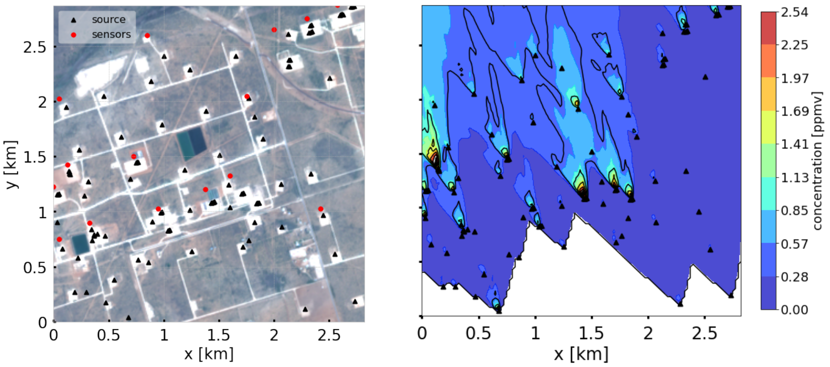
$$R_{mn^\star} = \arg\max_{\mathcal{P}^s} \mathcal{P}^s_m \, R^s_{mn^\star}. \tag{8}$$

Finally, to each predicted ranking we can assign a probability obtained by multiplying the (selected) marginal posterior point estimate of the source and the predictive likelihood: $P(\boldsymbol{q}^\star_{s^\star}, \boldsymbol{\theta}|\boldsymbol{w}) \simeq P(\boldsymbol{w}|\boldsymbol{q}^\star_{s^\star}, \boldsymbol{\theta}) \, P(\boldsymbol{q}^\star_{s^\star})$.

The end-to-end source attribution process is re-assumed in the flow chart of Fig. 1. In the next section, we apply the methods described here to a scenario of practical relevance.

3 Case Study

The scenario we considered is of direct practical relevance as it can be prohibitively expensive to monitor large areas of interest with a 1:1 or higher sensor-to-source ratio. IoT sensors transmit real-time data on methane concentration and weather readings. An anomaly detection algorithm is employed to detect abnormal methane emissions; if anomalous readings are detected, these are flagged to the source attribution system that returns the most likely location(s) of the leak. The setup of the experimental AoI is shown in Fig. 2. We consider an AoI of approximately 9 km^2, in the Permian Basin in West Texas and Southeastern New Mexico, for our study. The Permian Basin is one of the most prolific oil and gas basins in the US, and contains numerous oil and gas infrastructure assets, many of which are likely emitters of methane. The scenario that we consider for our study is one where 100 possible sources are monitored by 15 high resolution methane sensors in the AoI. Sensor locations have been determined using the sensor placement optimization procedure detailed in Wang [25]; the optimization output is shown in Fig. 2, where sensors are placed either close to ground level or at heights of 5 and 10 m. In the next sections, we discuss data collection and processing. We will also describe how the test scenario and sensor readings were simulated in the absence of field sensor data.



(a)

(b)

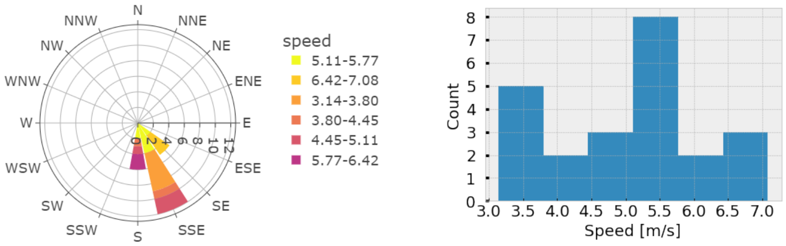
Fig. 2: a) Aerial view of the Area of Interest, showing locations of emission sources (black triangles) and sensors (red dots). b) Methane concentration (above background level) map in ppmv, on 20-07, at 3 pm. Level curve represent z direction.

3.1 Data collection and processing

We gather various inputs required for the Bayesian analysis in the AoI, such as weather variables, historical methane leak rate data, and oil and gas facility maps. We obtain hourly weather data (wind speed, wind direction, temperature, pressure, cloud coverage) from the weather station closest to the study area, from the National Oceanic and Atmospheric Administration (NOAA) Integrated Surface Dataset (ISD) [1]. The wind rose diagram for our AoI is shown in Fig.3a for a given test date and time. Methane emissions data can be obtained from aerial surveys or IoT sensor measurements or from historical knowledge of leaks from specific oil and gas assets. For our study, we leverage data from an extensive airborne campaign across the Permian Basin from September to November of 2019 [7] that quantified strong methane point source emissions (super emitters) at facility-scales. Since this data corresponds to leak rates from super emitters, we have a tunable parameter to scale the leak rates down. For our analysis, we scale it down by a factor of 3 to better represent the order of magnitude of leaks from normal methane emitters, while maintaining the heavy-tailed distribution shape from the original Permian Basin airborne campaign data set. We find the data to be in good agreement with an exponential distribution. Oil and gas facilities locations data, including wells, natural gas pipelines and processing plants, available in the public domain, are ingested for the area of interest [3, 22]. Satellite map of the AoI is obtained from Sentinel-2 data [2].

3.2 Scenario Simulation

Given the input data defined in the previous section, we use the forward model to simulate the methane concentration at each sensor location. The simulation is

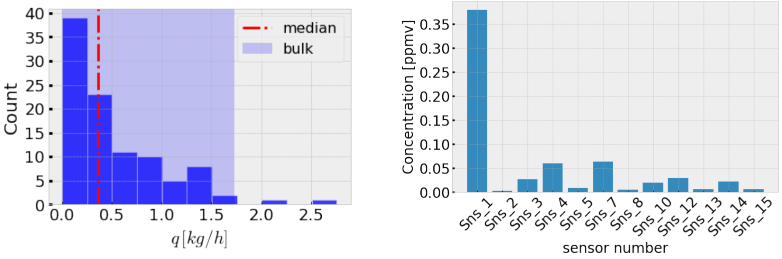


(a) Wind rose.

(b) Wind speed distribution.

Fig. 3: Wind speed and angle distribution for a specific day (07-20-2020).

performed on an area of approximately 9 km^2, and up to 200 m in the z-direction; the grid size (dx, dy, dz) is (25, 25, 5) m. The number of sources is 100, and the number of sensors is 15 (see Fig.2). We sample the leak rate of each source from the fitted exponential leak rate distribution and use weather data at the time of detection as an input to the plume model defined in Sect. 2.2. As there are no interaction terms in Eq. 2, the concentration field at each point is assumed to be additive. As a consequence, the compound signal at a sensor location is evaluated via summation of individual source contributions. Finally, Gaussian noise is applied to the readings, with a standard deviation corresponding to the sensor's systematic error, together with a detection threshold; for both parameters, we have used values reported by the sensor's vendor of 0.002 ± 0.0001 ppmv over background level (estimated at 1.8 ppmv in the AoI). Throughout this paper, we always report concentration values over background. In Fig. 4a we show an example input data for the leak rates; this shows a typical pattern where most sources have low emissions with few of them being anomalous, i.e. outliers. This is one of the most challenging scenarios we encountered, and we present it here in detail; in practice, there are many possible scenarios, the most favourable being when all the sensors can capture a strong signal. We comment on these other results at the end of Sect. 3.3. In Fig. 4a we use Tukey's fence criteria to separate the bulk of the sample from the outliers; the shaded area in the plot is determined by the interval $[Q_1 - \alpha\,\mathrm{IQR}, Q_3 + \alpha\,\mathrm{IQR}]$, where $\mathrm{IQR} = Q_3 - Q_1$ is the inter quantile range, and Q_1, Q_3 the quantiles. A value of $\alpha = 1.5$ is used to determine the outliers, while $\alpha = 3$ determines extreme values. There are three outliers, corresponding to sources $[10, 80, 92]$, with source 10 being the highest emitter. The median separates low (50%) from average (47%) emitters, with high leak rate outliers constituting only the remaining 3%. However, not all emissions are measured by the sensors, as the concentration value depends on both weather conditions (determining dispersion) and sensor positioning. For the example considered here, one of the high leak rate outliers (80) is not captured at all by the sensors. The compound sensor readings are then used as an input to the source attribution algorithm as detailed in the next section.



(a) Leak rates sample.

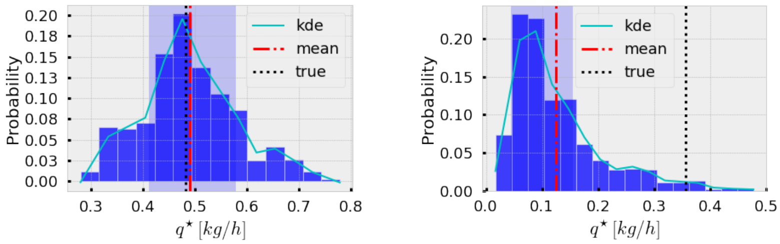
(b) Sensor measurements.

Fig. 4: a) Sample from the leak rate distribution, together with median and bulk; see text for details. b) Simulated measurement at sensor location on 07-20-2020 at 3 pm. This sample shows a typical pattern where most sources have low emissions, with few of them being super-emitters. In this scenario, only 12 of the deployed 15 sensors report above threshold readings.

3.3 Results

We leverage the `Pyro` [4] implementation of the No-U-Turn, Hamiltonian Monte Carlo [9] to sample the marginal posterior; the entire process is summarized in Fig. 1. We found that a relatively small collection of 1000 samples provides a good compromise between accuracy and computational time; the sampler returns the leak rate distribution for each of the 100 sources, at different degrees of convergence. As we are using priors obtained from empirical data, these are not necessarily conjugated, hence the marginal posterior distribution is unknown and needs to be fitted. Although it is possible to look for a continuous parametric fit, here we opt to use the Kernel Density Estimation (KDE) implementation in `scikit-learn` [4] using grid search with cross validation to fix the kernel and bandwidth of each leak rate distribution. In Fig. 5 we show two examples of distributions where convergence is achieved and where it is not. Each figure shows the histogram of the samples, the KDE fit, the 68% Highest Posterior Density Interval (HPDI), together with two vertical lines showing true value and sample average. Following the discussion of Sect. 2.3, we use 51 sample point estimates within the HPDI, for each leak rates marginal posterior distributions. In general, we found the posterior sample average to be a robust central estimator; in addition we use 50 percentiles points estimate (from 0 to 100 at steps of two). The point estimates are used in the forward model to estimate the predictive likelihood and the source contribution at each sensor. The latter are used in the ranking model to extract the top three sources, per sensor, contributing the most to the measured methane concentration, together with their ranking confidence. After this process, we are left with 51 ranking and concentration values for each sensor. In the final step, for each sensor, we select the the maximum (predictive) likelihood value out of the 51 evaluated and use this as our best estimate for

[4] `https://scikit-learn.org/`



(a) Source 92.

(b) Source 9.

Fig. 5: Marginal posterior distributions: samples histogram and KDE fit are shown together with a 68% HPDI interval (shaded area), the sample mean and the true value. In Fig. 5a, the sample mean provides a good estimation of the the true value, while this is not the case in Fig. 5b, where it lies in the tail of the distribution.

likely sources. We can visualize this result via the network map in Fig. 6, showing sensor to source connectivity for the three selected sources, weighted by source leak rate. Ultimately, this constitutes the model recommendation presented to the monitoring operator, to help them plan further field investigation and plan leak remediation by prioritizing the most likely source of leakage. For testing, we evaluate the mean average precision [19] at $k = 3$ (mAP@3); as our intent is to detect the highest emitting sources at each sensor, $k = 3$ represents a good compromise between keeping this focus while looking at mid-level emissions as well. As we explained later, the performance of the model decreases when including more sources, as optimization samples are dominated by the higher emitters. mAP@3 evaluates how many of the three proposed sources have been correctly ranked, and average the result over all available sensors. For the example above, we find mAP@3=0.86. Crucially, the ranking error depends on the relative magnitude of the source's leak rates, this being true also for the regression metrics presented in the next section. Fig. 7 shows the true and predicted source contributions to the methane concentration signal and leak rates detected at a sensor.

Leak rate quantification We have repeated the same analysis for 10 more days randomly sampled through the year at different times of the day. Depending on factors such as the weather, leak rate sample and crucially the number of sensors recording the signal (as low as 1), the mAP@3 may vary, although on average is still ~ 0.83, showing the robustness of the model. Leak rate quantification and source attribution are both outputs from the Bayesian learning algorithm. Accurate leak size estimation is critical in quantifying the environmental footprint of methane leaks, and is also crucial from a regulatory and governance perspective, helping companies build trust with stakeholders and the public. Fig. 7b shows the true and predicted leak rate estimates for the highest 3 emitters (sources) whose signal is detected at a sensor. We use Mean Absolute Percentage Error

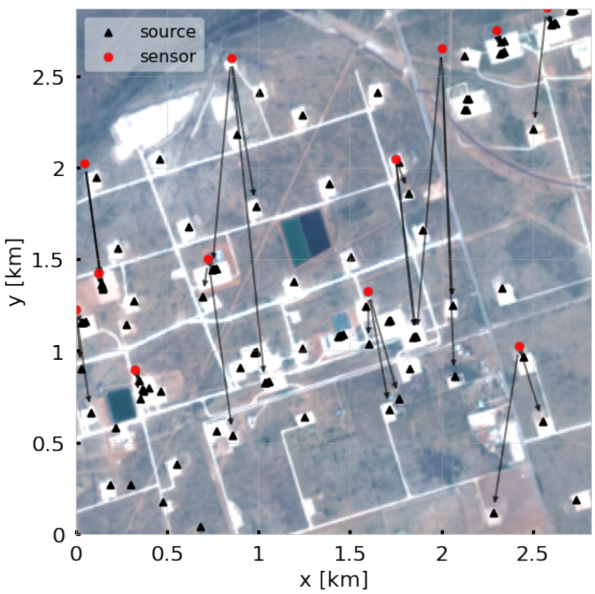
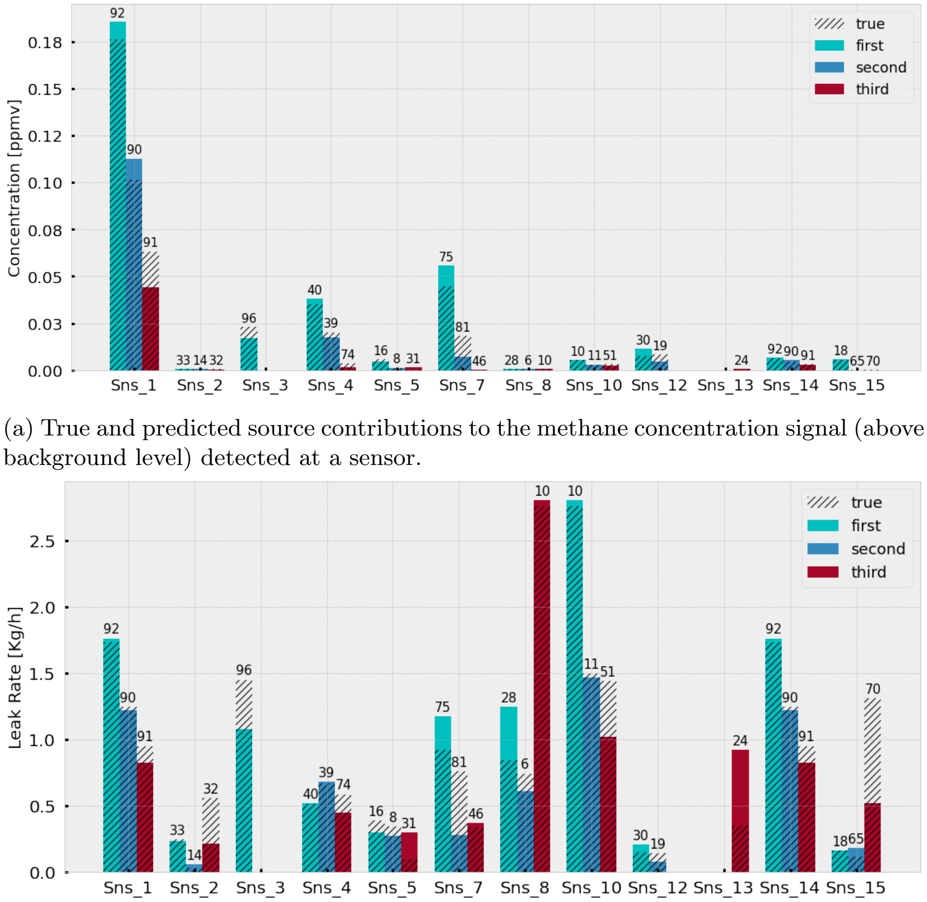


Fig. 6: Network Map: Connections between sensors and sources are used to visualize attribution

(MAPE) [5] as the metric to evaluate the performance of the leak rate quantification algorithm. When evaluated for sources that have been correctly classified as contributing to the signal at a sensor (as determined by the mAP@3 metric), we obtain a total MAPE $\simeq 29\%$. As we mentioned in the previous section, routine sources are more difficult to estimate due to their lower leak rates and the skew nature of the distribution; following Fig. 4a and the discussion of Sect. 3.2, when breaking down the error into low, medium and high leak rates (outliers) we find the corresponding MAPE to be: 50%, 24% and 1.7%, showing how medium and high leak rates can be reliably estimated. We found this behaviour to be consistent across different different scenarios, see Sect. 3.3. This leak rate estimate can also be used to update the prior leak rate distribution, which can then be used for future analyses. This can be thought of as a Bayesian learning process that iteratively improves the source attribution and estimation process.

4 Conclusions

We have presented a Bayesian source attribution and quantification model applied to the realistic situation of non linear dependence between concentration and leak rates, and a regime where the number of sources to monitor greatly exceeds the number of ground field sensors, mathematically corresponding to an under-determined system. We use the mean average precision at $k = 3$ (mAP@3)



(a) True and predicted source contributions to the methane concentration signal (above background level) detected at a sensor.

(b) True and predicted leak rate estimates for the highest 3 emitters (sources) whose signal is detected at a sensor.

Fig. 7: The number on top of each bar represents the source ID. Only correctly classified sources are shown.

for evaluating the performance of the source attribution algorithm, and observe a mAP@3=0.86 for the experiments performed, which signifies that 86% of leaks detected at sensors were correctly attributed to the true sources; we found this result to be robust across different weather and leak rates sample scenarios, with an average mAP@3 ~ 0.83. For leak rate quantification, we use MAPE to evaluate model performance, and we report a total MAPE of 29%. Breaking down this error by the relative size of the leak rates, we find that most of the estimation error comes from low emitting sources, obtaining a MAPE of 24% and 1.7% for medium and high emitters, respectively. The leak rate quantification for sources with high signal strength at sensors is significantly more accurate than that for sources with relatively lower signal strength. Accurate leak source attribution and quantification are vital for any methane Leak Detection and Remediation program, and for addressing regulatory and governance aspects, where an accurate assessment of the environmental impact of such leaks is critical.

5 Future Work

Spatial heterogeneity in weather data and transient plume behavior have a crucial impact on the atmospheric dispersion of methane; these effects are not captured by the simple Gaussian plume model used in this work. When choosing the forward model, one needs to balance precision vs. computational time. In this respect, the use of modern machine learning methods to approximate complex modeling constitutes a promising way forward. Some of these methods not only allow us to replace physics-based solvers, but also to learn directly from a mix of real and simulated data. In this work we have also restricted our analysis to very small sample sizes (1000) when performing Bayesian optimization; the choice is due to favouring response time vs. higher accuracy, the former being the most important factor in deployment. We are exploring the use of Stochastic Variational Inference as a replacement for the more costly Hamiltonian Monte Carlo together with more informative likelihood distributions. Finally, access to sensor data will allow us to better estimate model parameters, including a more realistic account of total noise, beyond the systematic error currently modeled.

References

1. NOAA ISD datasets. https://www.ncei.noaa.gov/products/land-based-station/automated-surface-weather-observing-systems/
2. Sentinel-2 imagery. https://sentinel.esa.int/web/sentinel/missions/sentinel-2?msclkid=7c80fe6cc7ba11ec9c5499250f796bd7
3. Texas Railroad Commission datasets. https://www.rrc.texas.gov/resource-center/research/data-sets-available-for-download/, accessed: 2021-09-13
4. Bingham, E., Chen, J.P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P., Goodman, N.D.: Pyro: Deep universal probabilistic programming. The Journal of Machine Learning Research **20**(1), 973–978 (2019)
5. Bowerman, B.L., O'Connell, R.T., Koehler, A.B.: Forecasting, time series and regression: An applied approach. South-Western Pub (2005)
6. de Coninck, H., Revi, A., Babiker, M., Bertoldi, P., Buckeridge, M., Cartwright, A., Dong, W., Ford, J., Fuss, S., Hourcade, J.C., et al.: Strengthening and implementing the global response. In: Global warming of 1.5 C: Summary for policy makers, pp. 313–443. IPCC-The Intergovernmental Panel on Climate Change (2018)
7. Cusworth, D.H., Duren, R.M., Thorpe, A.K., Olson-Duvall, W., Heckler, J., Chapman, J.W., Eastwood, M.L., Helmlinger, M.C., Green, R.O., Asner, G.P., et al.: Intermittency of large methane emitters in the permian basin. Environmental Science & Technology Letters **8**(7), 567–573 (2021)

8. García, J.G., Hosseini, B., Stockie, J.M.: Simultaneous model calibration and source inversion in atmospheric dispersion models. Pure and Applied Geophysics **178**(3), 757–776 (2021)
9. Hoffman, M.D., Gelman, A., et al.: The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. J. Mach. Learn. Res. **15**(1), 1593–1623 (2014)
10. IEA: Driving down methane leaks from the oil and gas industry – technology report (Jan 2021), https://www.iea.org/reports/driving-down-methane-leaks-from-the-oil-and-gas-industry
11. Jeong, H.J., Kim, E.H., Suh, K.S., Hwang, W.T., Han, M.H., Lee, H.K.: Determination of the source rate released into the environment from a nuclear power plant. Radiation protection dosimetry **113**(3), 308–313 (2005)
12. Klise, K.A., Nicholson, B.L., Laird, C.D.: Sensor placement optimization using chama. Tech. rep., Sandia National Lab.(SNL-NM), Albuquerque, NM (United States) (2017)
13. Lushi, E., Stockie, J.M.: An inverse gaussian plume approach for estimating atmospheric pollutant emissions from multiple point sources. Atmospheric Environment **44**(8), 1097–1107 (2010)
14. McKain, K., Wofsy, S.C., Nehrkorn, T., Stephens, B.B.: Assessment of ground-based atmospheric observations for verification of greenhouse gas emissions from an urban region. PNAS Earth, Atmospheric and Planetary Sciences **109:22**, 8423–8428 (1912)
15. Panofsky, H.A., Prasad, B.: Similarity Theories and Diffusion. Int. J. Air Wat. Poll. **9**, 419–430 (1965)
16. Richardson, L.F.: Atmospheric diffusion shown on a distance-neighbour graph. Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character **110**(756), 709–737 (1926)
17. Rogelj, J., Shindell, D., Jiang, K., Fifita, S., Forster, P., Ginzburg, V., Handa, C., Kheshgi, H., Kobayashi, S., Kriegler, E., et al.: Mitigation pathways compatible with 1.5 c in the context of sustainable development. In: Global warming of 1.5 C, pp. 93–174. Intergovernmental Panel on Climate Change (2018)
18. Rudd, A., Robins, A.G., Lepley, J.J., Belcher, S.E.: An inverse method for determining source characteristics for emergency response applications. Boundary-layer meteorology **144**(1), 1–20 (2012)
19. Salton, G., J, M.M.: Introduction to modern information retrieval. McGraw-Hill (1983)
20. Stockie, J.M.: The mathematics of atmospheric dispersion modeling. Siam Review **53**(2), 349–372 (2011)
21. Taylor, G.I.: Diffusion by continuous movements. Proceedings of the London Mathematical Society **2**(1), 196–212 (1922)
22. U.S. Energy Information Administration: Layer information for interactive state maps (2020), https://www.eia.gov/maps/layer_info-m.php
23. Veigele, V.J., Head, J.H.: Derivation of the Gaussian Plume Model. Journal of the Air Pollution Control Association **28:11**, 1139–1140 (1978)
24. Wade, D., Senocak, I.: Stochastic reconstruction of multiple source atmospheric contaminant dispersion events. Atmospheric Environment **74**, 45–51 (2013)
25. Wang, S., Malvar, S., Nunes, L., Whitehall, K., Oruganti, Y.D., Alaudah, Y., Badam, A.: Unsupervised machine learning framework for sensor placement optimization: analyzing methane leaks. In: NeurIPS 2021 Workshop on Tackling Climate Change with Machine Learning (2021), https://www.climatechange.ai/papers/neurips2021/70