PHYSICS-GUIDED INFERENCE FROM GAS EMISSION DATA FOR SOURCE CHARACTERIZATION AND AIR POLLUTION MAPPING

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Amir Montazeri

December 2021

PHYSICS-GUIDED INFERENCE FROM GAS EMISSION DATA FOR SOURCE CHARACTERIZATION AND AIR POLLUTION MAPPING

Amir Montazeri, Ph.D.

Cornell University 2021

Releases of pollutants into the atmosphere pose risks to human health, the environment, and the economy. Fine-scale monitoring of air pollution and efforts to characterize the impact of the local environment on pollutant concentrations have led to great developments in sensor technology and the generation of a staggering amount of new data. While the advent of large datasets has led to more emphasis on statistical modeling, the disconnectedness of these models from underlying physical laws degrades their generalizability. Consequently, the objective of the present research is to apply a combination of statistical and deterministic models to quantify the effects of the environment at varying scales on: 1) our ability to observe and measure pollutant concentration profiles, 2) our ability to make inferences about the state of the pollutant sources, and 3) The evolution of pollutant concentration profiles.

This dissertation is divided into three parts. The first part focuses on a theoretical analysis of the uncertainties involved in leak quantification via gas imaging techniques. These uncertainties are quantified through statistical analysis of Large Eddy Simulation (LES) data. Our results show that uncertainties that are due to inferring the 3D plume structure from 2D projections become smaller as measurements are made at larger downwind distances from the emission source. Further, acquisition times on the order of tens of seconds are sufficient to significantly reduce these uncertainties.

The second part employs a recursive Bayesian scheme to infer the varying states of a gas emission source observed through downwind mobile measurements. Our findings suggest that the statistics of the measurements, such as the coefficient of variation and range are good predictors of the performance of the Bayesian algorithm. In addition, the algorithm shows a high success rate in detecting the state change when the emission rate is tripled.

The third part introduces a spatial clustering framework developed for studying the role of land-use in mediating the effect of meteorology on urban air quality. Our study is based on long-term mobile measurements of Nitrogen Dioxide (NO_2) concentrations in Oakland, California. We find strong correlations between wind speed and NO_2 concentrations in the absence of other causes of strong vertical mixing such as highway traffic and higher surface heat fluxes in summer. In addition, an analysis of the exceedance probabilities shows that wind speed is effective in lowering the highest concentrations even in regions where mean concentrations are not responsive to wind speed. These findings coupled with projections of climate (e.g., wind speed forecasts) and urban development can be used to make predictions regarding future air quality in urban areas.

BIOGRAPHICAL SKETCH

Amir Montazeri was born and raised in Shiraz, Iran where he attended Shahid Dastgheib high school, a branch of the National Organization for Development of Exceptional Talents (NODET). He ranked 15 in the National Entrance Exam of Universities in Iran out of more than 270,000 participants before moving to Cambridge, UK in 2011 to study engineering. Amir received his M.Eng degree in Aerospace engineering from University of Cambridge with his award winning final project on Aerodynamics and acoustics of a NASA open rotor. In 2015, he enrolled as a Ph.D. student in the Sibley school of Mechanical and Aerospace Engineering at Cornell University where he worked on the inverse problem of source inference from gas emission observations.

To my parents and my brothers, Nima and Sina

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my advisor, John Albertson for his valuable guidance during my time at Cornell, and for being an endless source of ideas. I would like to thank Edwin (Todd) Cowen for serving on my committee, teaching one of the best courses that one can take at Cornell in "Experimental Methods in Fluid Dynamics", and having me as a teaching assistant which was a great learning experience for me. I would also like to thank Gregory Bewley for serving on my committee and his helpful teaching and feedback along the way. I am also grateful to Peter Diamessis for granting my first teaching assistantship at Cornell during which his valuable advice helped me become a better teacher and a better researcher. Special thanks to Xiaochi (Joe) Zhou, for his mentorship during my first three years as a graduate student and collaboration on most of the work presented in this thesis. I also would like to thank Achim Lilienthal for collaboration on the land use clustering study and his thorough review of my manuscript.

Many thanks to my fellow graduate students at the Environmental Fluid Mechanics & Hydrology (EFMH) group, Theo, Xiao, Dan, Katie, Gerardo, Pierre, Nidia, Gustavo and Yanle for their support and feedback on my seminar presentations. I am especially thankful to Xiao for his helpful comments on statistical methods and to Theo for being a great colleague in our semesters as TAs.

Beyond my collaborators, mentors and colleagues, I would not have been able to reach the finish line without the support of friends in Ithaca and beyond. First, I like to thank my housemates over the years. Rahmtin, for sharing a passion for cooking elaborate Persian dishes and complex video games, and numerous conversations on all aspects of life. Jonathan, for being the most effective communicator and the most relaxed housemate. Steffen, for teaching

me a lot about Western cooking through his love for Pasta. Molly, for being a dependable friend with whom I had many thoughtful conversations about life. Mischa, for stimulating conversations on American politics and way of life and excellent taste in music. Aditya, for sharing numerous youtube videos that brightened my day and conversations about life after graduate school.

During my time as a graduate student, I spent numerous hours on video calls with friends who supported me while being far away, whom I like to thank. Pouya, who has been my closest friend since we were 10. His insight into academia, industry, work and life has started many ideas and conversations that we discuss to this day and I believe has enriched both our lives. Ali, for nerdy conversations on topics that only the two of us find interesting, from small details in a movie to high school math and physics problems. Oddly enough, our other discussions revolve around trivial subjects like the impacts of climate change, the latest scientific discoveries and the meaning of life.

I also like to thank my partner, Anya for her love and kindness and for tolerating my endless rants. I believe that spending time with her has made me a better, kinder and happier person.

Last but not least, I would like to thank my family. My parents, whom I miss dearly, for all their sacrifices throughout the years. Without their trust in my decisions and unconditional support, I would have not been able to pursue my goals and dreams. Nima, for always being someone I can depend on and being an amazing role model. Sina, for always believing in me, even when I was consumed by self-doubt. His constant support from afar kept me sane during the cold winters in Ithaca. Finally, my cousin Soodeh, who helped me settle in the U.S. and gave me invaluable advice on how to navigate life as a graduate student.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Releases of pollutants into the atmosphere pose risks to human health, the environment, and the economy. Efforts for successful sensing of these pollutants at fine scales and characterizing the impact of the local environment on their evolution have led to great developments in sensor technology. Examples of these new sensor technologies include but are not limited to hyperspectral imaging systems, networks of point sensors, networks of line sensors, personal and portable air pollution sensors, and network of mobile sensors. This explosion in sensor technology has led to the generation of a staggering amount of new data and thanks to advancements in wireless connectivity and cloud computing, constructing statistical models is easier than ever. Therefore, application of machine learning and big data approaches has gained a lot of traction in recent years. However, the lack of reliance of these approaches on underlying physical laws and unavailability of proper training data sets for some applications, degrades their performance in practice.

Traditionally, the inference of the state or dynamics of physical systems has relied on either physics/equation based (deterministic) or data-driven (statistical) models. Effective utilization of deterministic models requires good specification of parameter values in addition to an outstanding understanding of the underlying physics of the system. When building deterministic models of complex systems, parameters of the system components are often unavailable because of incomplete technical specifications, hidden physical interactions, or interactions that are too complex to model from first principles [4]. As a result, we often resort to simplifying assumptions and coarser models that imperfectly

describe the behavior of the system, which not only leads to a poor description of system behavior but also renders the model difficult to comprehend and analyze [5]. For example, the Air Quality Dispersion Model (AERMOD) that is the dispersion modeling system recommended by the U.S. Environmental Protection Agency (USEPA), receives regular updates for capability improvement. In particular, recent analyses related to modeling building downwash have shown AERMOD to both overpredict and underpredict ground-level concentrations in the building wake, depending on the building dimensions; stack height; stack location; and the orientation of the building relative to the wind direction [6,7].

Purely data-driven models on the other hand, ignore any knowledge of the underlying physics of the system and often require large amounts of labeled training data to successfully model the behavior of the system under varying conditions. There are two primary hurdles for successful modeling of physical systems through purely data-driven methods. First, complex systems involve a large number of variables that follow intricate and non-stationary patterns that dynamically change over time. Data-driven models for these systems often suffer from scarcity of representative training examples that cover the entire domain of variables. Therefore, the true nature of relationships between the variables cannot be established through the limited number of labeled instances. In particular, data-driven models can often generate spurious relationships that fit the training data well, but do not generalize well outside the available data [5]. Second, the black-box nature of data-driven methods leads to models that lack interpretability and faithfulness to the underlying physical laws (e.g., conservation of mass). As a result, these models have limited predictive power especially when it comes to extreme events with no previous record. Consequently, a prudent approach is to use methods that use the physics of the system and prior

knowledge about the domain to guide the construction of data-driven models.

The objective of the present research is to apply a combination of statistical and deterministic models to quantify the effects of the environment at varying scales on: 1) The evolution of pollutant concentration profiles, 2) our ability to observe and measure pollutant concentration profiles, and 3) our ability to make inferences about the state of the pollutant sources. The impacts of the environment on the evolution of concentrations of pollutants can be attributed to the meteorology and the spatial domain in which the pollutants are released in. These impacts indicate that air pollution is an area that is sensitive to projected climate change and anthropogenic developments. For example, a few studies have observed correlations of high-ozone events (>80 ppb) with meteorological variables such as temperature, regional stagnation and wind speed among others [1,8,9]. Figure 1.1 illustrates the observed probability that the maximum daily 8 hour average ozone will exceed 80 ppb based on the daily maximum temperature for 3 regions in the United States. The figures shows that in the Northeast, the probability can double for a 3K increase in temperature, highlighting the potentially large sensitivity to climate change.

Recent investigations into climate and anthropogenic influences on air quality have revealed many challenges. In terms of the effects of meteorology on pollutant concentrations, these challenges arise mainly because some meteorological parameters are heavily correlated, making it difficult to separate the effect of individual variables on pollutant levels. In addition, the meteorological effects vary based on the spatial domain. In terms of anthropogenic developments, the wide variety of spatial environments makes it challenging to develop universal models to quantify effects of the spatial domain on air quality. While

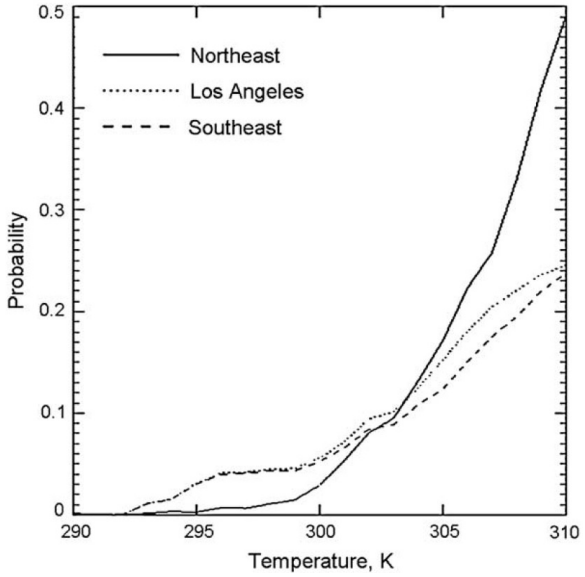

Figure 1.1: Observed probability that the maximum daily 8-h average ozone will exceed 80 ppb for a given daily maximum temperature, based on 1980-1998 data. Values are shown for the Northeast U.S., the Los Angeles Basin, and the Southeast U.S. [1].

some of these challenges (e.g., correlation between meteorological parameters) can be overcome by gathering more data, it is infeasible to collect data under every possible scenario (e.g., it is impossible to gather air pollution data for every possible building arrangement). Meanwhile, deterministic models in this context are often based on several assumptions that are rarely satisfied in complex practical environments, and in some cases, they are difficult to verify. Thus, we propose approaches that combine statistical and deterministic models with the goal of characterizing the state and dynamics of environmental variables that drive air pollution.

Accomplishing the goal of the present research is consequential in the effort to mitigate air pollutant emissions and make predictions regarding extreme events caused by changes to the local controls. To this end, the research is divided into three parts presented in chapters 2-4.

In chapter two, the viability of hyperspectral imaging systems for quantifying emissions is investigated. Based on the cost of the imaging systems compared to more traditional leak quantification methods, a leak rate estimation skill to within 20% error is desired. The threshold of 20% is chosen based on the challenge set by ARPA-E's Methane Observation Networks with Innovative Technology to Obtain Reduction (MONITOR) program [10]. The viability of the imaging systems is quantified in a theoretical setting where we present a systematic categorization of the involved uncertainties with a focus on projection uncertainties that arise when 3D velocity and concentration fields inside a plume are projected on 2D images. The projection uncertainties are then quantified using Large Eddy Simulation experiments of a point source release into the atmosphere. Our results show that projection uncertainties are higher when the gas plume observations are made at smaller distances to the point source. Further, acquisition times on the order of tens of seconds are sufficient to significantly reduce the projection uncertainties. We use these findings alongside practical consideration to suggest guidelines for more robust leak quantification through gas imaging.

Quantification of gas emissions (especially methane) is an important and active area of research that is necessary for mitigating emissions, computing national inventories, and understanding of the sources and magnitudes of emissions from various industries. However, in recent years, researchers have noticed "fat-tail" distributions of methane emissions in the oil and gas industry, suggesting that a small fraction of sites are responsible for the majority of emissions [2,3,11]. For example, Figure 1.2 shows the distribution of emissions from both active and idle well pads sampled in California as measured by a recent mobile monitoring campaign, indicating that 15% of the sampled active well

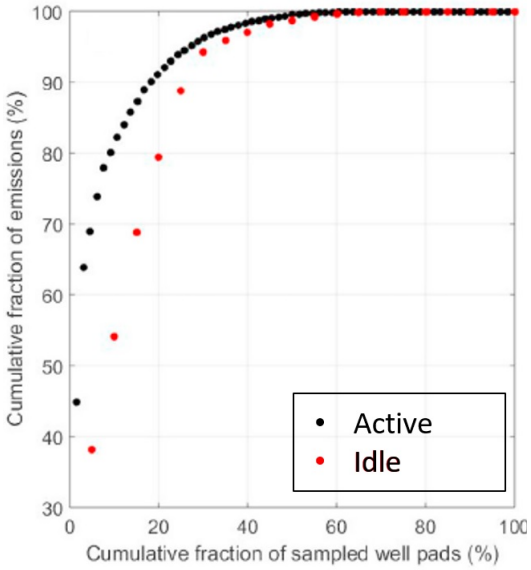

Figure 1.2: Cumulative fraction of measured emissions as a function of the cumulative fraction of the sampled well pads for active and idle well pads in California [2]

pads are responsible for more than 85% of the total measured emissions. Similar results have been observed across different sectors of the oil and gas industry in various regions in the United States and Canada [3, 12].

Discrepancies between bottom-up and top-down estimates of emissions has led to the belief that the biggest emitters (often referred to as super-emitters) arise due to abnormal operating conditions [3, 11]. Here, top-down refers to aggregate emission estimates across large geographies through usage of aircraft, satellites or tower networks, while bottom-up methods aggregate and extrapolate emissions from individual equipment made directly at the emission point. Figure 1.3 shows the difference between bottom-up and top-down estimates for sites in the Barnett Shale production region in Texas, where 10,000 Monte-Carlo simulations were performed to estimate the range of emission estimates from bottom-up measurements. In particular, the highest emitting 1% of sites in the

observed top-down distribution have cumulative emissions nine times larger than the bottom-up estimates. With bottom-up estimates often underestimating the effects of abnormal operations, it is therefore concluded that these operating conditions cause super-emitters. Therefore, rapid and effective identification of faulty operations can lead to substantial emission mitigation. To put the effect of rapid identification of faulty operations into perspective, let us consider an annual inspection scheme that would detect any unwanted leaks at a production site. Under the assumption that the leak can occur at any time between inspections, with the initiation times of leaks evenly distributed throughout the year, the average fault would persist for $365/2 = 182.5$ days. Consequently, a fault identification scheme that could detect leaks within a few days would lead to substantial emission reductions (e.g. finding the leak within 18 days results in a 90% emission reduction on average).

In chapter three, we present a changepoint detection algorithm based on a recursive Bayesian scheme that allows for simultaneous emission rate estimation and fault detection. Our Bayesian inference methodology is superior to current methods for estimation and fault detection [13], as it readily yields the uncertainties related to emission estimates and allows for ”online” fault detection (i.e. fault detection with more measurements becoming available incrementally). The proposed algorithm is tested on a series of near-field controlled release mobile experiments, with promising results demonstrating successful detection (>90% success rate) of changes in the leak rate when the emission rate is tripled after an abrupt change. This result is significant, because faulty operations often lead to emission rates that are orders of magnitude larger than pre-fault rates. Moreover, we show that the statistics of the measurements, such as the coefficient of variation and range are good predictors of the performance of

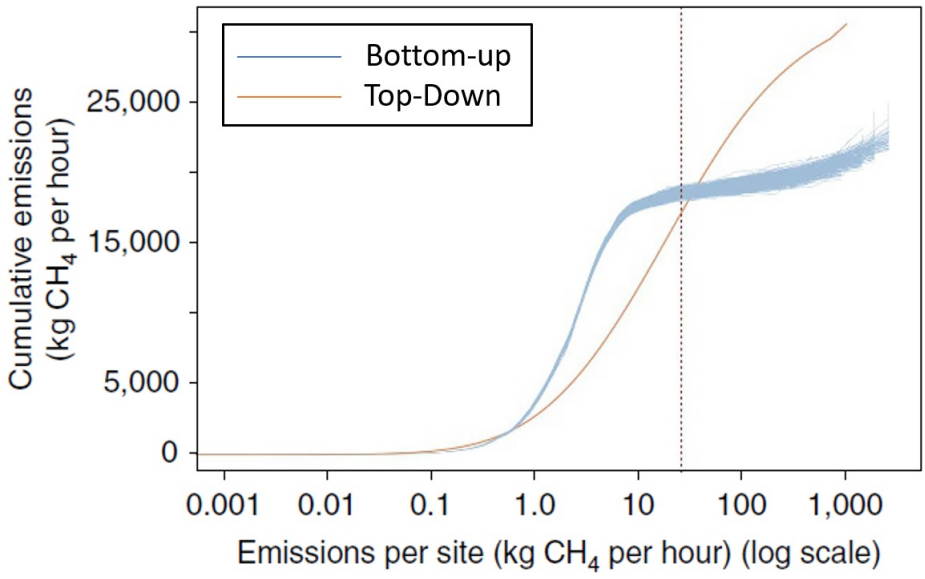

Figure 1.3: Cumulative emissions as a function of emission rate per site. Blue lines represent each of 10,000 Monte Carlo iterations from the bottom-up aggregation reported, orange line represent the top-down results, vertical line represent the 99th percentile of site emissions [3].

the algorithm. Finally, we describe how this methodology can be easily adapted to suit time-averaged concentration data measured by stationary sensors, thus showcasing its flexibility.

While the focus of chapter three is on targeted small scale mobile measurements and their applicability for emission rate estimation and fault detection, chapter four relies on large scale mobile measurement campaigns that cover large spatial domains (such as entire urban areas) over long periods of time (i.e. months and years).

In chapter four, we examine the role of urban land use in mediating the effect of regional meteorology on Nitrogen Dioxide (NO_2) concentrations measured by mobile monitors in different regions of Oakland, CA. Inspired by land-use

Regression (LUR) models, we cluster 30-meter road segments in the urban area based on their land use. The concentration data from the resulting clusters are stratified based on seasonality and conditionally averaged based on concurrent wind speeds. The clustering analysis yielded 7 clusters, with 4 of them chosen for further statistical analysis due to their large sample sizes. Two of the four clusters demonstrated in winter a strong negative linear relationship between NO_2 concentration and wind speed. A weaker correlation and flatter slope was found for the cluster representing road segments belonging to interstate highways. No significant relationship was found during the summer season. These findings are consistent with the concept of strong vertical mixing due to highway traffic and increased surface heat fluxes during summer weakening the relationship between wind speed and NO_2 concentrations. In summary, the clustering analysis framework presented here provides a novel tool for use with large-scale mobile measurements to reveal the effect of urban land form on the temporal dynamics of pollutant concentrations and ultimately human exposure.

CHAPTER 2

ON THE VIABILITY OF VIDEO IMAGING IN LEAK RATE QUANTIFICATION: A THEORETICAL ERROR ANALYSIS

2.1 Introduction

With technological advances in extraction techniques [14], the production of Natural Gas (NG) in the United States underwent a steady increase in the 2010s and reached a new record high in 2019 [15]. Operational and accidental emissions at NG production, processing and transmission facilities release methane, the major component of NG, into the atmosphere, posing risks associated with climate change and health and safety [16]. Therefore, mitigation of emissions has become a top priority in the United States, highlighted by the introduction of periodic leak detection and repair (LDAR) surveys for methane by the U.S. Environmental Protection Agency (EPA) in the 2016 updates to the New Source Performance Standards [17]. In practice, LDAR programs following U.S. EPA's Method 21 or using optical gas imaging (OGI) are effective for component-level leak detection, however, they are labor and resource intensive, which prevents frequent survey and prompt mitigation efforts. While the focus of LDAR programs has been on leak detection and localization, quantification of emission sources can lead to more effective emission mitigation by prioritizing repair for larger leaks. Furthermore, quantification of small and medium-sized emission sources adds depth to our understanding of emission profiles associated with the NG supply chain that can lead to further mitigation efforts.

In recent years, new measurement systems and technologies have been developed to quantify emissions of methane from equipment and operations with

reduced cost and/or improved spatial coverage. Few examples include portable methane analyzers [18], open-path laser spectrometers [19], remote sensing of methane from aircraft and satellite [20–22], and ground-based mobile sensing approaches [2, 23]. Use of portable methane analyzers is an accurate method for quantifying known emission sources, but localization efforts are often labor intensive with the requirement of investigating entire facilities at a slow pace. Open-path laser spectrometers are used to quantify emissions from facilities, but require equipment for wind measurement in addition to knowledge of location of emitting components that necessitates accompanied use of detection methods such as OGI. Satellite and airborne methods allow coverage of large areas and detection of relatively large emission sources (e.g. detection limit for the Airborne Visible/Infrared Imaging Spectrometer - Next Generation is 240 kg CH_4/day) [22]. However, such high detection limits prevent identification of low- and moderate-emission sources under typical meteorological conditions. Ground-based mobile sensing approaches are useful in quantifying emissions from facilities without offering solutions for component leak localization [24, 25].

One newly developed technique that allows for quantification and localization is ground-based remote sensing via gas imaging cameras [26–28]. Gas imaging includes capturing video images of methane plumes in the environment to quantify emission rates. Briefly, this technique involves the comparison of the at-sensor radiant energy in the IR part of the electromagnetic spectrum in the presence and absence of the methane gas plume. This difference in radiance is then related to the depth integrated concentration of the gas (also known as the concentration-path length) in ppm×m through the use of Beer-Lambert law and the temperature contrast between the gas and the background

scene. This technique shows promise as it offers high spatiotemporal resolution in mapping gas concentrations as well as possibility of automation and continuous monitoring of sites. In addition, ancillary equipment for wind measurement is not required, since with high frequency imaging (e.g. >1Hz) gas velocities inside the plume can be approximated by tracking plume features in consecutive images using velocimetry algorithms such as minimum quadratic differences [29], cross-correlation between consecutive images [30], and block-matching [26]. With the measured methane concentration and the estimated flow velocity, the emission rates can be computed based on the principle of mass conservation. However, a systematic analysis focusing on the uncertainty of such estimates has been lacking in the literature.

The effectiveness of gas imaging techniques in quantifying unknown leak rates is tied to the level of uncertainty in leak quantification. Furthermore, lower uncertainties in leak rate quantification lead to lower false alarm rates and promote effective mitigation of emissions and reduced costs [12]. An in-depth understanding of the sources of uncertainty is essential for increasing the accuracy and precision of leak quantification. To this end, we divide the uncertainties into the following categories: 1) instrumentation, 2) operational and 3) two-dimensional (2D) projection uncertainties. Lower uncertainties can generally be achieved through technological advances in equipment and instrumentation. For example, increased spatial and temporal resolution of imaging cameras translate into lower uncertainty in concentration measurement and velocity estimation which in turn leads to lower total uncertainty in leak quantification [12]. Instrumentation uncertainties are usually reported by manufacturers and have been previously studied in detail [26, 27, 30, 31]. Meanwhile, the transformation of concentrations and velocities into emission rates is also an un-

certain process. The uncertainties of this transformation process can be due to operating conditions (operational uncertainty) or arise from approximating the three-dimensional (3D) methane plume, using a 2D view as seen by the camera (projection uncertainties). Operating conditions such as distance between camera and leak, background temperature, changes in wind speed and direction can affect uncertainty levels by affecting detection capabilities of the cameras and have been previously investigated for leak detection through simulations and experiments [32,33]. On the other hand, projection uncertainties have been largely ignored in the literature. It is worth noting that while operating conditions and instrumentation uncertainties can affect the magnitude of projection uncertainties, projection uncertainties are ubiquitous irrespective of other uncertainties. In other words, even with perfect equipment and algorithms accurately measuring concentrations and velocities, projection uncertainties will still be present.

In this chapter, we formulate the projection uncertainties related to the transformation of gas imaging measurements into emission rates through a rigorous theoretical analysis that couples mass balance and spatial Reynolds decomposition. Our analysis is first carried out through comparison of a 3D view of instantaneous plume transport and its 2D projection which models an emission scene as observed through the lens of a camera. Our analysis divides the projection uncertainties into two distinct uncertainty expressions. These two expressions are then quantified and compared against each other using Large eddy simulations (LES) of a point source plume dispersion under neutral atmospheric conditions. Furthermore, the effects of acquisition time and downwind distance from leak on projection uncertainties are quantified. Finally, we discuss the implications of these results on the viability of gas imaging techniques

on leak rate quantification.

It is worth noting that this chapter focuses on fugitive emissions of methane due to the recent incentives for accurate quantification, namely the DOE ARPA-E's Methane Observation Networks with Innovative Technology to Obtain Reduction (MONITOR) program [10], however, our theoretical analysis stands true for video observations of any release of a conserved scalar into the environment.

2.2 Theory

To evaluate the projection uncertainties in leak rate quantification through gas imaging, we need to derive expressions for the error terms involved. To this end, in section 2.2.1 we describe an instantaneous view of plume transport and use it to formulate an exact solution to a point source leak rate problem. This formulation has been previously used to characterize point sources through mobile sensor data with details available in [34]. In section 2.2.2, we introduce a 2D projection of the transport formulation to model images captured in gas imaging experiments. We use these 2D projections to formulate an approximate solution to the leak rate problem. The difference between the exact and approximate solutions explicitly describes the projection uncertainties present in leak rate quantification as shown in section 2.2.3.

2.2.1 Instantaneous 3D view of plume transport

We consider the release of a gas (e.g. methane) from a point source into the environment. The release is happening in the surface-layer of the atmospheric boundary layer (ABL) where the assumptions of statistically stationary and horizontally homogeneous turbulence hold [35]. Without loss of generality, we assume that the $x-$axis of the coordinate system is directed along the mean wind, and we refer to the y and z directions as "depth" and "height" directions, respectively. This setup is shown in Figure 2.1a, with the point source located at (x_p, y_p, z_p). In this setup, u, v and w are defined as velocity components in x, y and z directions, respectively. A control volume is defined starting from the origin O and extending in the three principal directions up to downwind sampling positions x_m, from y_{min} to y_{max} in depth, and z_{min} to z_{max} vertically. The control volume contains the source and is defined such that the plume generated from the point source only exits the face on the $y-z$ plane at $x = x_m$. Conservation of mass states that the source rate (mass per time), Q can be expressed as

$$Q = F(x_m, t) + \frac{dS(t)}{dt}, \tag{2.1}$$

where $S(t)$ is total mass of the emitted gas in the control volume, t is time, and $F(x_m, t)$ is the mass flow rate out of the control volume.

The total mass of the emitted gas in the control volume at any time, t, is calculated by integrating the above-ambient concentration over the full control volume as follows

$$S(t) = \int_0^{x_m} \int_{y_{min}}^{y_{max}} \int_{z_{min}}^{z_{max}} c(x, y, z, t) dz dy dx, \tag{2.2}$$

where c is the above-ambient gas concentration. In the ABL, the flow is highly turbulent so that molecular diffusion can be ignored relative to turbulent trans-

port [36]. Therefore, the mass flow rate exiting the downwind face of the control volume is related to the gas concentration and velocity as

$$F(x_m, t) = \int_{L_z} \int_{L_y} c(x_m, y, z, t) u(x_m, y, z, t) dy dz, \tag{2.3}$$

where L_y and L_z are the plume depth and height, respectively.

Equations (2.1), (2.2) and (2.3) allow for quantification of the source rate with knowledge of the gas concentration and plume velocity as functions of space and time. These expressions are considered the benchmark to which we will compare other formulations to evaluate their intrinsic uncertainties. Note that the precise local gas concentrations and velocities are not readily available through gas imaging techniques; therefore, we simulate the image sampling process in a 2D model framework so as to quantify the truncation errors introduced by the measurements and analytics.

2.2.2 2D modeling of instantaneous plume transport

In order to model the images of plume transport, the control volume introduced above is projected such that the $y-$axis is collapsed and only the x and z principal directions are resolved (Figure 2.1b). In this scenario, it is not possible for a camera to obtain the velocity and concentration variations with depth (i.e. variations in $y-$direction). Instead, a depth-integrated concentration profile is observed [27,28], which is denoted by c^y and defined as

$$c^y(x, z, t) = \int_{L_y} c(x, y, z, t) dy. \tag{2.4}$$

Equation (2.4) can be utilized to show that in the 2D model the total mass of the emitted gas within the control volume $S(t)$ can be evaluated exactly.

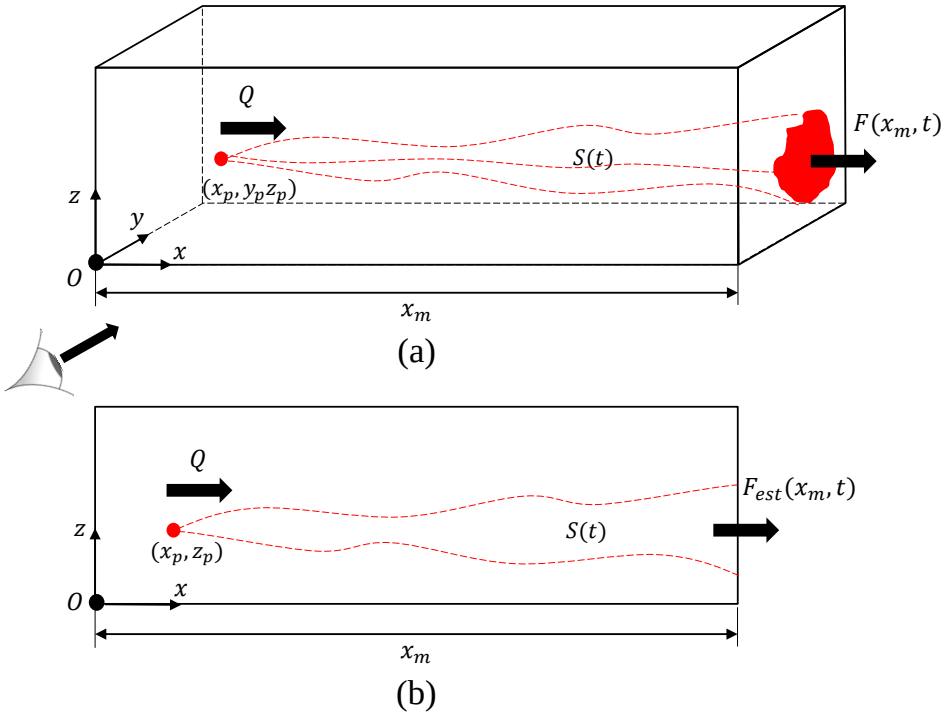

Figure 2.1: A control volume containing a continuous source with a mass flow rate of Q located at (x_p, y_p, z_p) in (a) a 3D, with a cross-plane view of the plume mass flow rate, $F(x_m, t)$ at downwind distance, x_m and (b) a two-dimensional snapshot modeling an image obtained via gas imaging leading to an estimate of the mass flow rate, $F_{est}(x_m, t)$ at downwind distance, x_m.

To calculate the mass flow out of the control volume, the velocities inside the plume are also required. In practice, these velocities are measured by employing optical velocimetry algorithms [26, 29, 30] that track the depth-integrated concentration profiles over consecutive images. The mass flow out of the control volume can be estimated using these *inferred velocity profiles* from plume tracking, labeled $u_i(x_m, z, t)$, and the depth integrated concentration profiles as follows

$$F_{est}(x_m, t) = \int_{L_z} c^y(x_m, z, t) u_i(x_m, z, t) dz, \tag{2.5}$$

where F_{est} is the estimated outward mass flow. A detailed discussion on the possibilities for the inferred velocity is presented in section 2.4.2.

2.2.3 Projection uncertainty formulation

The projection uncertainty associated with the 2D projection of the plume can be formulated by comparing the leak rate quantification procedures of sections 2.2.1 and 2.2.2. We continue the analysis under the assumption that instrumentation uncertainties are negligible, meaning that the depth-integrated concentrations measured through gas imaging are without significant error. We note that although such an assumption is not valid in practice, it allows us to isolate and estimate the projection uncertainties. In practical applications, the projection uncertainties should be added to other uncertainty estimates for a better quantification of the total uncertainties. With this consideration, $S(t)$ can be calculated exactly through the 2D measurement inference algorithm. Therefore, the projection uncertainties in quantifying the leak rate are solely dependent on the difference between F and F_{est}.

The relationship between F and F_{est} can be written explicitly by applying Reynolds decomposition to the dependent variables (u and c) and decomposing them into a spatial mean and a fluctuating part, e.g. for the velocity, $u(x, y, z, t) = \overline{u}(x, z, t) + u'(x, y, z, t)$. Here, $\overline{u}$ denotes the depth-averaged velocity measured across the depth of the plume, and u' is the corresponding fluctuating velocity. This decomposition directly leads to the following results

$$\int_{L_y} u'(x, y, z, t)dy = 0, \tag{2.6}$$

$$\int_{L_y} c'(x, y, z, t)dy = 0, \tag{2.7}$$

$$\int_{L_y} \overline{c}(x,\ z,\ t)\ dy = c^y(x, z, t). \tag{2.8}$$

For simplicity of notation, x, y, z, t will be dropped for the remainder of this

section.

By utilizing Reynolds decomposition, a relationship between F and F_{est} can be rigorously derived. First, we apply the decomposition to u and c to rewrite F as follows

$$\begin{aligned} F(x_m,t) &= \int_{L_z}\int_{L_y} cudydz = \int_{L_z}\int_{L_y} (\overline{c}+c')(\overline{u}+u')dydz \\ &= \int_{L_z}\int_{L_y} \left(\overline{c}.\overline{u} + \overline{c}u' + \overline{u}c' + c'u'\right) dydz \\ &= \int_{L_z} \overline{u}c^y dz + \int_{L_z}\int_{L_y} c'u'dydz, \end{aligned} \tag{2.9}$$

where equations (2.6)-(2.8) are used to simplify terms along the way. Subtracting equation (2.5) from equation (2.9) yields the difference between F and F_{est}

$$F(x_m,t) - F_{est}(x_m,t) = \int_{L_z} (\overline{u}-u_i)c^y dz + \int_{L_z}\int_{L_y} c'u'dydz. \tag{2.10}$$

With the assumption that instrumentation uncertainties are negligible, the right hand side of equation (2.10) describes the projection uncertainties present in leak rate quantification via gas imaging, since they are caused by using a 2D projection of the plume to approximate the mass flow rate. The first integral in equation (2.10) scales with the difference between the true depth-averaged velocity and the inferred velocity estimate from the 2D image analysis. Therefore, prediction of the scale of this velocity difference under typical application conditions indicates the importance of the first term. The second integral describes the covariance of velocity and concentration fluctuations. Hereafter, we will refer to the first integral as the ”mean velocity error term” and the second integral will be referred to as the ”covariance error term”. We explore the scale of these terms by analysing a dataset acquired through Large Eddy Simulations (LES), as described in the following sections.

2.3 Large Eddy Simulation data

Large Eddy Simulation is used to create a virtual test site for simulation of the dispersion of a passive scalar (e.g. methane) in the surface layer of the ABL. The LES turbulent modeling is particularly useful for simulating high-Reynolds number flows in the ABL. The LES code used in this study has been utilized and validated in numerous studies [37–41]. In brief, the code numerically solves the resolved Navier-Stokes and mass conservation equations on a Cartesian grid while the unresolved (sub-grid) dynamics are closed in terms of the resolved scales [38].

LES has been previously used as a realistic proxy for the space-time evolution of plumes in turbulent near-neutral environments [41–44]. Therefore, in this study, methane release from a point source is simulated under near-neutral turbulent conditions in an unobstructed flat homogeneous terrain.

The virtual site was set up with 0.469 m horizontal and 0.188 m vertical grid resolution with a total simulation domain size of 60m in x (along-wind) and y (crosswind) directions and 15m in z (vertical) direction (128x128x80 spatial resolution). The virtual site was constructed to resemble the Methane Emissions Technology Evaluation Center (METEC) well pads, a facility funded through the ARPA-E's MONITOR program and built to provide a location that models natural gas production sites. In the virtual site, the source is located at a height of approximately 2.25m to match the average height of a typical leak as modeled in the METEC facilities. A 30-minute spin-up period was implemented to allow the simulated turbulence to reach a statistically stationary state. In this case, the average and standard deviations of the wind and scalars approach a

Table 2.1: Summary of parameters used in LES.

Name	**Value**
Computational domain size ($x_{max}, y_{max}, z_{max}$)	60, 60, 20 (m)
Computational grid size ($\Delta x, \Delta y, \Delta z$)	0.469, 0.469, 0.188 (m)
Height of source (z_p)	2.27 (m)
Sampling frequency (f_s)	1 (Hz)
Sampling duration (T_s)	900 (s)
Downwind distance of intersects (x_m)	4.7, 7.0, 8.9, 17.9, 26.8 (m)
Normalized downwind distance of intersects (x_m/z_p)	2.1, 3.1, 4.0, 8.0, 11.9 (-)

constant value [41]. For the analysis to follow, 5 $y-z$ intersects are created at non-dimensional downwind distances normalized by the source height $\left(x_m/z_p\right)$ of approximately 2, 3, 4, 8 and 12, on which the instantaneous velocities and concentrations are sampled (recorded) at a frequency of 1 Hz for the duration of 15 minutes. A summary of parameters used in the LES is presented in Table 2.1.

2.4 Results and Discussion

In this section, we utilize the LES dataset to estimate the scale of the covariance and mean velocity error terms as functions of sampling distance to leak and sampling duration.

2.4.1 Covariance error term

To observe the significance of the covariance error term compared to the mass flow and the leak rate, the LES data set is employed as follows. For each snap-

shot (saved at a frequency of 1 Hz), the covariance error term is calculated at the $y-z$ intersects located downwind of the emission source. Then, we compute the "normalized covariance error", denoted by Φ_c and defined as the ratio of the covariance error term to the leak rate

$$\Phi_c(x_m, t) \equiv \frac{\int_{L_z}\int_{L_y} c'u' dydz}{Q}. \tag{2.11}$$

Box plots of populations of Φ_c calculated at each of the 5 $y-z$ intersects distinguished by their downwind distance from the source, are presented in Figure 2.2. The figure shows that at all downwind distances from the source the covariance error term is almost always less than 10% of the leak rate. Further, the significance of the covariance error term drops as the downwind distance from the source is increased. A possible explanation for this result is through the assumption of local isotropy at the length scale of the plume. One consequence of local isotropy is the vanishing of all correlations between velocity components and scalars [45] leading to small values for Φ_c. Moreover, the drop in Φ_c with distance is due to the fact that at larger distances, concentration fluctuations from the mean become smaller as the plume widens, while the velocity fluctuations stay relatively constant. Meanwhile, the smaller than zero median and mean values suggest that the covariance error term is in a direction opposite to the total mass flow. This finding can be understood by noting that a higher velocity compared to the mean can move the plume and lead to lower local concentrations leading to observing opposite signs for u' and c' on average.

The tall whiskers in the box plots in Figure 2.2 indicate that in a single snapshot, Φ_c can take values in a relatively large interval. Therefore, it is plausible that mass flow rates be computed using multiple snapshots taken over a period of time, highlighting the effect of time-averaging on Φ_c. Figure 2.3 depicts the effect of time averaging on Φ_c for a normalized downwind distance of $x_m/z_p = 4$.

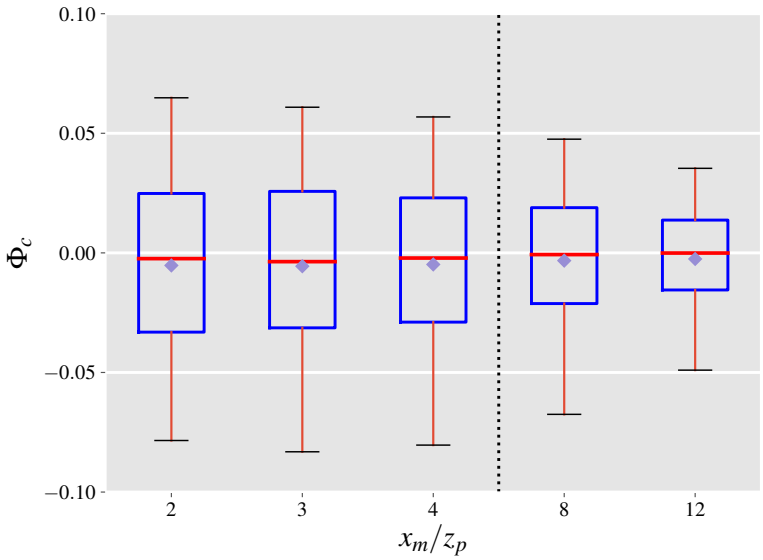

Figure 2.2: Distributions of the normalized covariance error term shown at 5 downwind $y-z$ intersects from the emission source measured for every saved snapshot from the LES. Box and whiskers plots show the median (red), 25th and 75th percentile (blue), the 5th and 95th percentile (black), and the mean (purple diamond) values of each distribution.

In this figure, the $y-$axis is labeled by $\langle\Phi_c\rangle$ to indicate a time-averaged parameter, where the angle brackets denote time-averaging. It can be seen that longer time-averages reduce Φ_c indicating that even short averaging times on the order of tens of seconds can lead to a substantial decrease in the significance of the covariance error term and therefore the projection uncertainty. It is worth noting that while averaging times longer than 30 seconds lead to further decreases in the normalized covariance error, they are not shown in Figure 2.3, because in practice, steady and statistically stationary wind conditions are uncommon for longer periods.

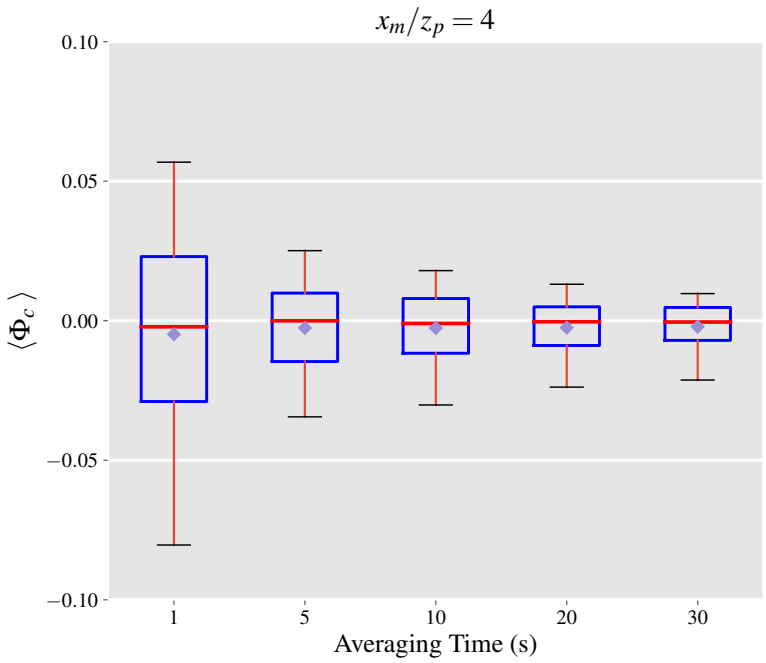

Figure 2.3: Effect of time-averaging on Φ_c at a normalized downwind distance of $x_m/z_p = 4$. Box and whiskers plots show the median (red), 25th and 75th percentile (blue), the 5th and 95th percentile (black), and the mean (purple diamond) values of each distribution.

2.4.2 Mean velocity error term

Similar to Φ_c, we define the "normalized mean velocity error" denoted by Φ_u, as the ratio of the mean velocity error term in equation (2.10) to the emission rate

$$\Phi_u(x_m, t) \equiv \frac{\int_{L_z} (\overline{u} - u_i)\, c^y dz}{Q}. \tag{2.12}$$

The normalized mean velocity error is therefore a function of the inferred velocity, u_i, which is dependent on the operational conditions and the velocimetry technique used to infer the velocity from gas imaging. There are numerous possibilities for defining the inferred velocities, among which we formally introduce and analyse three likely cases:

Case 1: Concentration weighted average velocity (ideal case)

In this ideal scenario, the inferred velocity is computed as a concentration weighted average velocity (also referred to as plume-weighted advection velocity) [34]:

$$u_i^{(1)}(x_m, z, t) \equiv \frac{\int_{L_y} ucdy}{\int_{L_y} cdy}, \tag{2.13}$$

where the superscript is used to show that these expressions are only valid for the corresponding case of the discussion regarding the mean velocity error term. Equation (2.13) can then be utilized to rewrite the mass flow out of the control volume

$$F(x_m, t) = \int_{L_z}\int_{L_y} cudydz = \int_{L_z} u_i^{(1)} \int_{L_y} cdydz = \int_{L_z} u_i^{(1)} c^y dz = F_{est}^{(1)}(x_m, t). \tag{2.14}$$

Therefore, employing the concentration weighted average velocity as the inferred velocity causes the mean velocity error term to cancel out the covariance error term in the projection uncertainty calculations, allowing for the mass flow out of the control volume to be computed exactly. In practice, this case in unlikely to be achieved, hence we continue by establishing an upper bound for the normalized mean velocity error.

Case 2: Maximum difference velocity (upper bound case)

To compute an upper bound for the normalized mean velocity error, the inferred velocity can be expressed as follows

$$u_i^{(2)}(x_m, z, t) \equiv \overline{u} + \max\left\{|u - \overline{u}| : y \in y_{plume}\right\}, \tag{2.15}$$

where y_{plume} corresponds to the interval of length L_y where the plume at (x_m, z) is instantaneously located. With this definition for the inferred velocity, we use

equation (2.12) to calculate the upper bound of the normalized mean velocity error denoted by Φ_u^u, with boxplots of populations at each of the $y-z$ intersects illustrated in Figure 2.4a. As the distance from the source is increased, the plume becomes wider, hence we expect the difference term, $\left(\overline{u}-u_i^{(2)}\right)$, to grow. However, as the plume grows wider through diffusion the depth-integrated concentrations at each height are decreased. As a result, Figure 2.4a shows that at closer distances to the source the rate of increase in Φ_u^u is faster compared to larger distances, with Φ_u^u almost staying constant between x_m/z_p of 8 and 12. The effect of time averaging on Φ_u^u is presented in Figure 2.5a highlighting that the range of uncertainties significantly drops as the averaging times are increased in a similar manner to the normalized covariance error. Further, the median value for Φ_u^u is under 0.20 which is a promising result for leak quantification using gas imaging based on current standards [10].

Case 3: Maximum concentration velocity

A plausible estimation for the inferred velocity is the velocity of a portion of the plume that has the highest concentration. In a gas flow, clumps of higher concentration contribute more to the measured depth-integrated concentrations than other parts of the plume. Therefore, it is expected that velocimetry techniques infer the gas velocity by tracking these highly concentrated clumps [31]. As a result, we define the inferred velocity in this case as follows

$$u_i^{(3)}(x_m,z,t) \equiv u\left(x_m, \arg\max_y c(x_m,y,z,t), z, t\right). \tag{2.16}$$

We use Φ_u^c to refer to the normalized mean velocity error computed with $u_i^{(3)}$ as the inferred velocity with boxplots of populations of Φ_u^c at 5 $y-z$ intersects illustrated in Figure 2.4b. In this scenario, the population distribution of

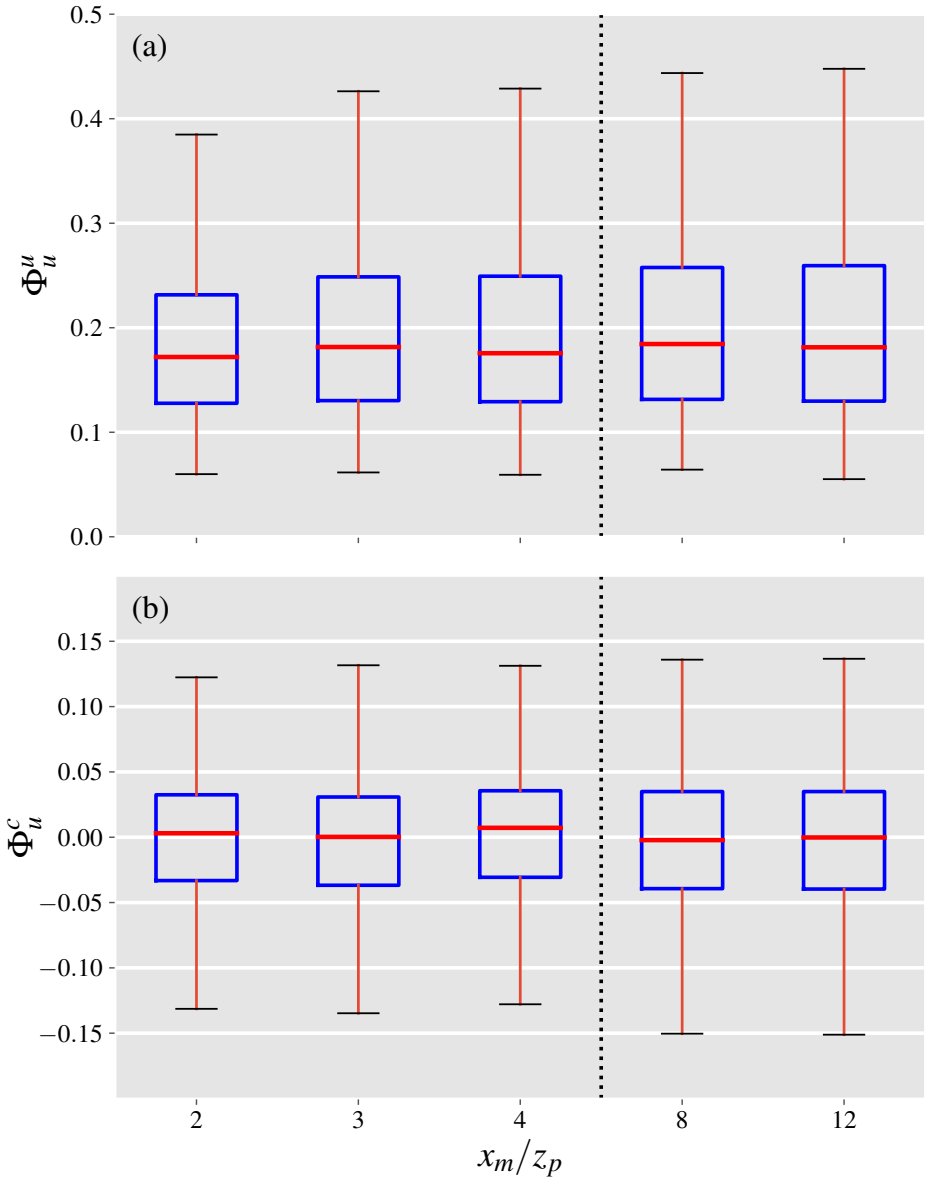

Figure 2.4: Distributions of (a) the upper bound of the normalized mean velocity error and (b) the normalized mean velocity error based on using the maximum concentration velocity as the inferred velocity at 5 downwind $y-z$ intersects from the emission source. Box and whiskers plots are as Figure 2.2.

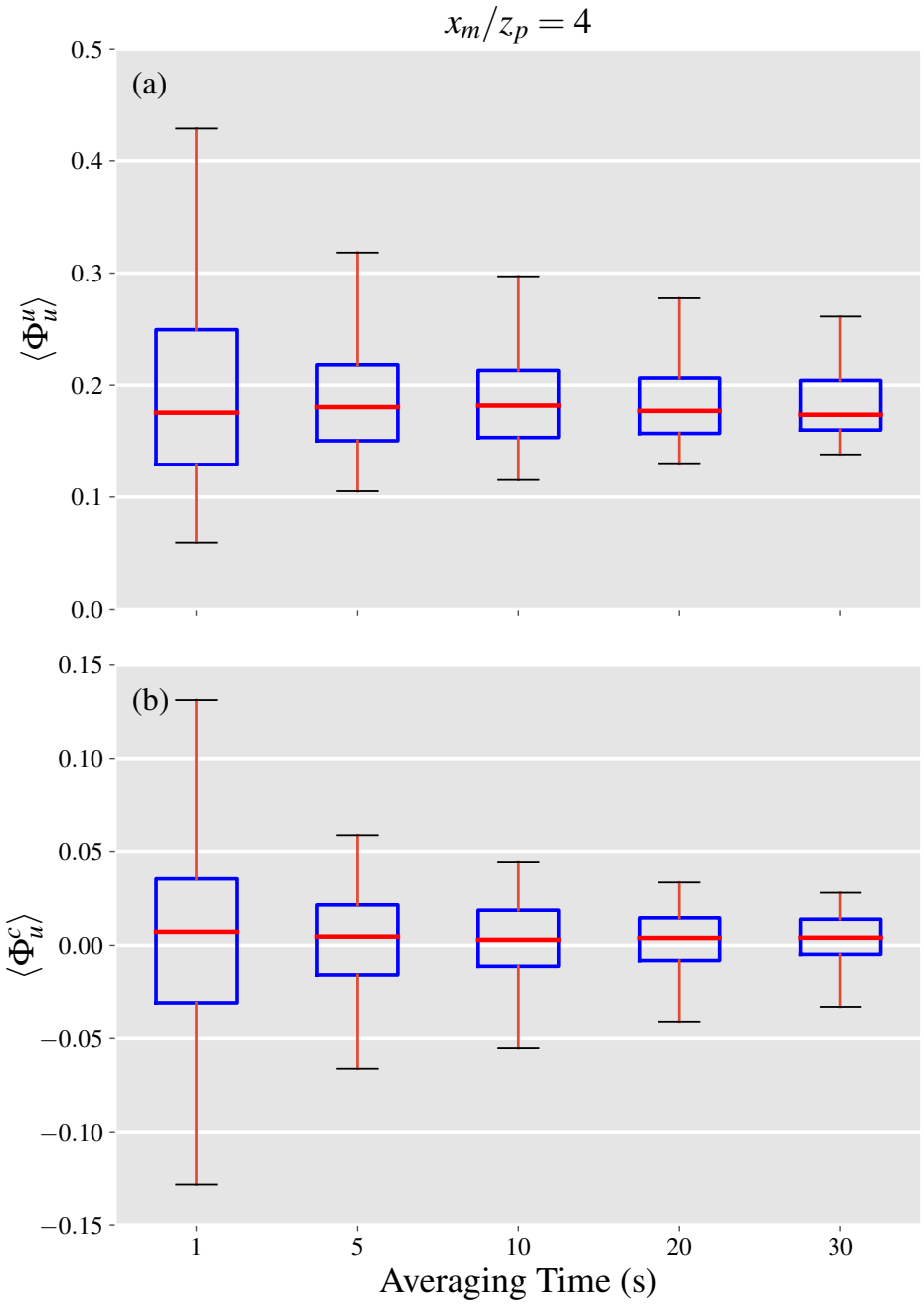

Figure 2.5: Effect of time-averaging on (a) Φ_u^u and (b) Φ_u^c at a normalized downwind distance of $x_m/z_p = 4$. Box and whiskers plots are as Figure 2.3.

Φ_u^c becomes wider with increasing distances from the source before reaching a plateau at normalized distances of 8 and 12 in a similar manner to Φ_u^u. For all considered distances, while the mean velocity error term can reach up to 15% of the leak rate in magnitude, the most likely scale of the error is between ±3 percent. Moreover, the median value of the distributions is within 1% of zero, suggesting that high acquisition times would result in minimal mean velocity errors. Figure 2.5b shows the effect of averaging time on the population distribution of Φ_u^c at $x_m/z_p = 4$, highlighting a significant drop of 50% in the width of the distribution (according to the 5-95 percentiles) when the averaging time is increased to 5 seconds. This finding alongside the effect of averaging times on the magnitude of the covariance error term, underlines the importance of longer acquisition times in order to reduce the projection uncertainties related to leak quantification. A more formal investigation of the total projection uncertainty is discussed in the next section.

2.4.3 Total projection uncertainty

Here, we define the normalized projection uncertainty as the ratio of the total projection uncertainty to the leak rate. By this definition, the normalized projection uncertainty ratio, denoted by Φ_t, can be computed by the addition of the normalized covariance and mean velocity errors. We use the inferred velocity from case (3) above to estimate Φ_t, with population boxplots depicted in Figure 2.6. The results indicate that the projection uncertainty drops as the outgoing surface of the control volume is constructed further away from the point source at x_m/z_p of 8 and 12. However, the reduction in the uncertainty range does not occur monotonically and the range only slightly varies between

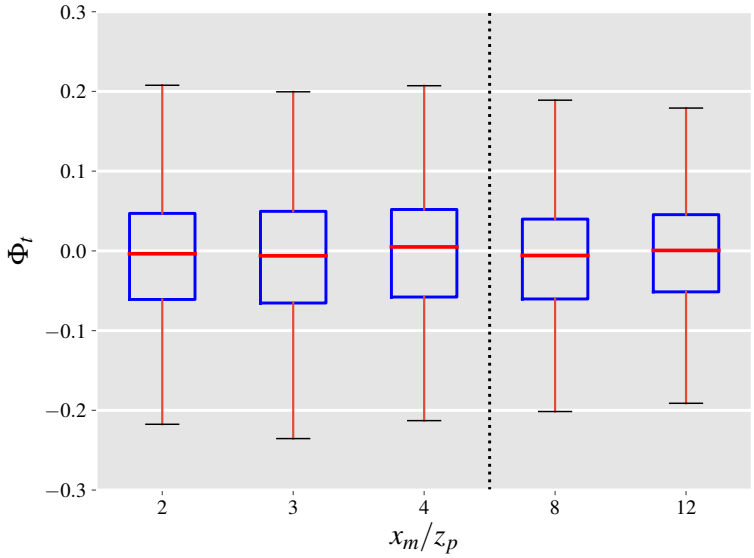

Figure 2.6: Distributions of the normalized projection uncertainty shown at 5 downwind $y-z$ intersects from the emission source. Box and whisker plots are as Figure 2.2.

normalized distances of 2, 3 and 4. Increasing the averaging time leads to narrower distributions of $\langle \Phi_t \rangle$ as presented in Figure 2.7 in a similar manner to the normalized covariance and mean velocity errors. A noteworthy observation is the diminishing returns of increasing the averaging time from 20 to 30 seconds with a relative drop of 17% (absolute drop of 0.02) compared to a relative drop of 54% (absolute drop of 0.22) when the averaging time is increased to from 1 to 5 seconds.

With the projected uncertainties being smaller at farther distances from the point source, it may seem desirable to observe gas plumes far downstream of the point source for the purpose of leak quantification (especially when long acquisition times are not possible). In practice however, the plume may be difficult to detect and quantify at large distances away from the source due to detection limits of the gas imaging instrument [32]. On the other hand, in addition

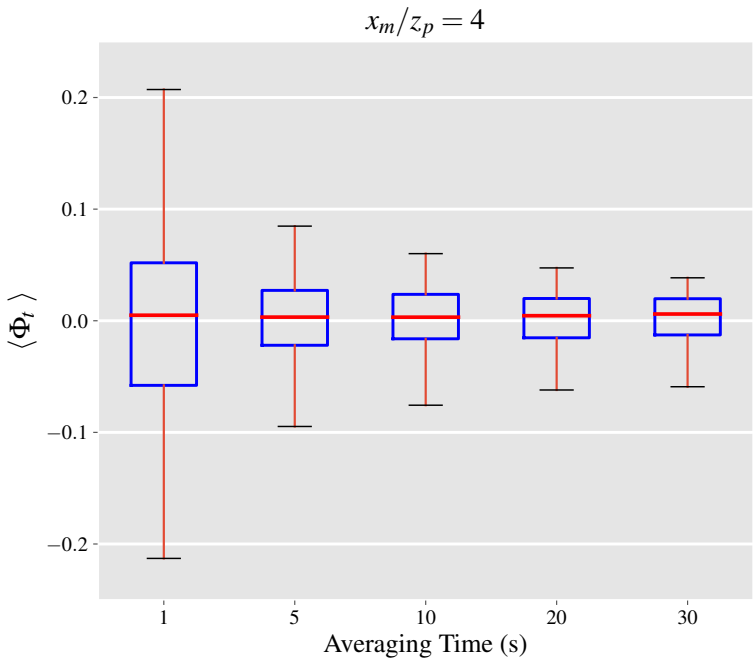

Figure 2.7: Effect of time-averaging on Φ_t at a normalized downwind distance of $x_m/z_p = 4$. Box and whisker plots are as Figure 2.3.

to higher projection uncertainties, imaging close to the point source can lead to underestimation if the saturation limit of the instrument in observing the depth-integrated concentrations is reached. Therefore, a plausible approach in leak rate estimation would be to employ several control volumes with control surfaces located at varying distances from the point source. The estimated leak rates from the control volumes can be averaged after removing the outliers created due to detection and saturation limits, to compute a final leak rate estimate. Moreover, the acquisition time should be as long as the wind conditions allow, since longer acquisition times are translated into lower projection errors.

2.5 Conclusions

In this chapter, we presented an approach for expressing and quantifying the projection uncertainties in estimating fugitive source emission rates through gas imaging techniques. The projection uncertainties arise from observing the dispersion of a 3D plume in the atmosphere through 2D video images. We developed a theoretical analysis that led to two separate terms associated with the projection uncertainties, the covariance and mean velocity error terms. A simulated dataset generated through Large eddy simulations was used to quantify the significance of each of the projection uncertainty terms under varying imaging constraints of acquisition time, and downwind distance from the leak source.

We found that low acquisition times and instantaneous estimates of the leak rate are prone to high projection uncertainties that can amount up to 20% of the emission rate. However, we expect the typical projection uncertainty to be between ±5% highlighting the potential of gas imaging techniques in leak quantification. In these cases, the covariance error term is responsible for between a quarter to a third of the projection uncertainties depending on the observed downwind distance from the leak source. Furthermore, we found that increasing the acquisition time by a few seconds can cause substantial (>50%) decreases in the projection uncertainties leading to much more robust estimates for the leak rate. The employment of long acquisition times and imaging far away from leak sources may prove difficult to achieve in real life releases, since long acquisition times require steady and statistically stationary wind directions that may be rare to occur in practical settings. Meanwhile, at farther distances downwind of leak sources the gas concentrations within the plume are likely

to drop below the detection limit of gas imaging cameras which can lead to underestimation of the leak rates. Consequently, for practical applications, we suggest the use of multiple control volumes at varying distances coupled with the longest acquisition times as allowed by the environmental conditions.

Altogether, the remote sensing approach based on the use of gas imaging technology is a promising technique that has the capacity for accurate leak quantification. This approach allows for non-intrusive leak quantification without the need for additional equipment for wind measurements. To the best of our knowledge, previous studies on leak quantification via gas imaging have only used a single control volume close to the source, whereas our findings suggest that estimates can be improved by employing multiple control volumes [26, 27]. With the development of new hyperspectral cameras recording images at high spatial and temporal resolutions and more efficient velocimetry algorithms, it is expected that the accuracy and speed in leak quantification can further improve as long as projection uncertainties are kept in check with our suggested guidelines.

CHAPTER 3

SIMULTANEOUS QUANTIFICATION AND CHANGEPOINT DETECTION OF POINT SOURCE GAS EMISSIONS USING RECURSIVE BAYESIAN INFERENCE

3.1 Introduction

Methane is a potent greenhouse gas (GHG) with a global warming potential (GWP) that is approximately 84 and 28 times greater than carbon dioxide (CO_2) on 20 and 100 year time scales, respectively [46]. Methane emissions from the oil and gas industry are among the largest anthropogenic sources of methane in the United States, accounting for approximately 30% of total emissions in 2019 [47]. Recent studies have found that emissions from almost all subsectors of the oil and gas supply chain demonstrate a "fat-tail" distribution, such that a relatively small number of large emitters are responsible the majority of emissions [2, 33, 48–52]. While the presence of these large emitters is concerning, it offers the potential of expedient reduction in GHG emissions and costs if the largest emitters are rapidly identified and repaired.

Measurements of methane emissions can be classified as either top-down or bottom-up [11]. Top-down studies rely on ambient methane measurements using aircraft, satellites, or tower networks to estimate aggregate emissions from all contributing sources across large geographies [11]. On the other hand, bottom-up methods aggregate and extrapolate emissions from individual pieces of equipment, operations, or facilities, using measurements made directly at the emission point or, in the case of facilities, directly downwind [11, 53–56]. Recent integrated research efforts have found that while emission estimates from

facility-based bottom-up approaches and top-down approaches are in agreement, these estimates are significantly higher than component-based estimates (i.e., when emissions are extrapolated from individual pieces of equipment) [3-4]. Detailed investigation of the discrepancy between component-based aggregates and estimates from other approaches (i.e., facility-based bottom-up and top-down methods) suggests that component-based methods miss high emissions caused by abnormal operating conditions (e.g., malfunctions). Such abnormal conditions are the defining attribute of large emitters that contribute the majority of emissions in the oil and gas sector. Therefore, prompt identification of abnormal process conditions (i.e., fault detection) can lead to substantial reductions in emission and costs for operators.

The abnormal operating conditions observed in the largest emitters are spatially and temporally variable [3, 22, 49, 57]. For example, emissions can significantly increase for a production site due to malfunctions at a certain point in time. Hence, emission reduction requires monitoring approaches that enable efficient and timely responses to the appearance of abnormal process conditions. Continuous monitoring offers the capability to rapidly detect faulty behavior that is necessary for reducing emissions from the largest emitters. Moreover, recent efforts to develop innovative technologies and algorithms to mitigate methane emissions have also highlighted the advantages of continuous monitoring over "snapshot-in-time" approaches in rapid identification of large emission sources [10, 58].

While continuous monitoring of oil and gas facilities are not commonplace yet, changes in the near future are likely due to the following: 1) Advancements in sensor technology and wireless communications allow for continuous mea-

surements to be made and stored in the cloud [58], and 2) monitoring mandates at the state and federal level, such as the U.S. Environmental Protection Agency (USEPA) fence-line monitoring program for early detection of benzene emissions [59] and continuous monitoring of air quality during pre-production and early-production of drilling operations producing gas and liquid hydrocarbons required by the Colorado Department of Public Health and environment [60]. Meanwhile, innovations in atmospheric inversion modeling are required to enable automatic emission estimation and fault detection using continuous measurements. It is worth noting that efficient and rapid fault detection can serve as an incentive for operators to employ continuous surveillance systems, as it can lead to significant reductions in unwanted emissions and costs.

Here, we propose a recursive Bayesian inference model that utilizes measurements from continuous surveillance systems (e.g., network of fixed sensors, and mobile sensors on unmanned vehicles) for simultaneous estimation and changepoint (fault) detection in point-source emission rates. Note that we use changepoint or fault to refer to a sudden increase in point-source emission rate as expected under abnormal operating conditions described earlier. The Bayesian inference model is useful in this context for multiple reasons: 1) It is equipped to deal with noisy data which in this case are caused by the stochastic nature of turbulence that drives the emitted gas and other measurement uncertainties, 2) it permits the determination of the uncertainty in estimated emission rates [61] and 3) it allows for "online" changepoint detection and emission estimates, i.e., the emission rate estimates and probability of detecting changes in emission rates are updated with every new measurement that arrives incrementally [62].

In this chapter, we first introduce an instantaneous plume dispersion formulation that is used to guide our Bayesian analysis. Then, the mathematical framework of the Bayesian inference and its application in point-source estimation and changepoint detection is described. Next, we apply the proposed Bayesian framework to near-field (with source-to-sensor distances $\leq$ 30m) mobile measurements of a controlled point-source emission and evaluate its performance in changepoint detection. Although the Bayesian framework and the plume transport theory in this chapter are tailored towards mobile measurements, they can be adapted for continuous measurements made by networks of fixed sensors.

3.2 Theory

Point-source characterization approaches often rely on time-averaged measurements. However, our analysis of the mobile sensor data requires a formulation for the instantaneous plume that is introduced in section 3.2.1. The presented formulation is applicable to passive scalars, which are diffusive contaminants in low concentrations such that they have no dynamical effect on the motion of the surrounding flowing fluid [63]. In subsequent sections, the following assumptions are made: 1) The emission rate from the point source is constant until an abrupt change causes the leak rate to increase to a new constant value. 2) The emission rates before and after the changepoint, and therefore mass concentrations before and after the changepoint are independent.

3.2.1 Instantaneous view of plume transport

Consider a steady-state point source located at the origin O of a local coordinate system (Figure 3.1). The wind velocity components in x, y, and z directions are defined as u, v, and w, respectively. We define a control volume starting at the origin O to a downwind vertical plane located at x_m, extending from y_{min} to y_{max} laterally, and from z_{min} to z_{max} vertically, such that the control volume encompasses the entire plume upwind of the mobile sensor. For this control volume, the application of conservation of mass yields the following expression for the emission rate (mass per time), Q_0:

$$Q_0 = F(x_m, t) + \frac{dS(t)}{dt}, \tag{3.1}$$

where $S(t)$ is total mass of the emitted gas in the control volume, t is time, and $F(x_m, t)$ is the mass flow rate exiting the control volume through the vertical plane at x_m. The control volume is defined such that no mass exits anywhere other than the vertical plane at x_m, therefore the mass flow rate can be expressed as

$$F(x_m, t) = \int_{z_{min}}^{z_{max}} \int_{y_{min}}^{y_{max}} c(x_m, y, z, t) u(x_m, y, z, t) dy dz, \tag{3.2}$$

where c is the mass concentration of the passive scalar. It is worth noting that in the atmospheric boundary layer, large Reynolds numbers are typically observed and the flow is highly turbulent, therefore, molecular diffusion is ignored [36].

It is useful to define a normalized distribution of the mass concentration, labeled D, and a plume-weighted advection velocity, labeled u_e, at the exit plane of the control volume as

$$D(x_m, y, z, t) = \frac{c(x_m, y, z, t)}{\int_{z_{min}}^{z_{max}} \int_{y_{min}}^{y_{max}} c(x_m, y, z, t) dy dz}. \tag{3.3}$$

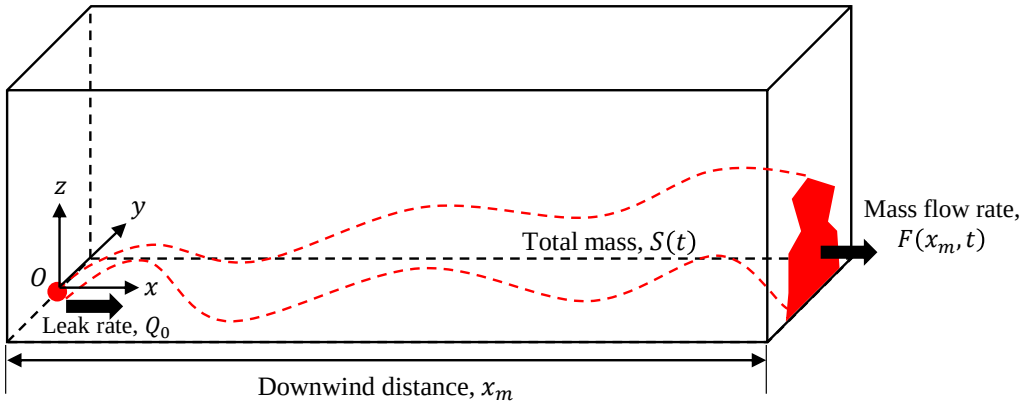

Figure 3.1: A control volume containing a point emission source (located at the origin, O) with a mass flow rate of Q_0, and a cross-plane view of the mass flow rate at downwind distance, x_m.

$$u_e(x_m, t) = \int_{z_{min}}^{z_{max}} \int_{y_{min}}^{y_{max}} D(x_m, y, z, t) u(x_m, y, z, t) dy dz, \tag{3.4}$$

We can rewrite $u_e(x_m, t)$ by substituting equation (3.3) into (3.4) and applying equation (3.2) as:

$$\begin{aligned} u_e(x_m, t) &= \frac{\int_{z_{min}}^{z_{max}} \int_{y_{min}}^{y_{max}} c(x_m, y, z, t) u(x_m, y, z, t) dy dz}{\int_{z_{min}}^{z_{max}} \int_{y_{min}}^{y_{max}} c(x_m, y, z, t) dy dz} \\ &= \frac{F(x_m, t)}{\int_{z_{min}}^{z_{max}} \int_{y_{min}}^{y_{max}} c(x_m, y, z, t) dy dz}. \end{aligned} \tag{3.5}$$

Equation (3.5) can then be used in conjunction with equation (3.1) to relate the mass concentration trajectory when traversing the plume, $c(x_m, y, z, t)$, to the other relevant variables as

$$c(x_m, y, z, t) = \frac{Q_0 - dS(t)/dt}{u_e(x_m, t)} D(x_m, y, z, t). \tag{3.6}$$

In practice, $u_e(x_m, t)$ can be approximated using nearby meteorological measurements. The vertical scaling of the wind profile based on the Monin-Obukhov similarity theory (MOST) can then be applied to adjust these meteorological measurements by height difference as needed [64], as detailed in

Appendix A.1. Accordingly, $u_e(x_m, t)$ is replaced by $u_e^M(x_m, t)\delta_u(x_m, t)$, where $\delta_u(x_m, t)$ accounts for the ratio between the actual and approximated plume-weighted advection velocities, $u_e(x_m, t)$ and $u_e^M(x_m, t)$, respectively. The superscript M is used to highlight model estimated quantities. Similarly, we introduce $\delta_S(x_m, t) = 1 - \frac{dS(t)/dt}{Q_0}$ to represent the non-steadiness in the total mass stored in the control volume, normalized by Q_0. Therefore, equation 3.6 can be expressed as follows

$$c(x_m, y, z, t) = \frac{Q_0}{u_e^M(x_m, t)} \left(\frac{\delta_S(x_m, t)}{\delta_u(x_m, t)} \right) D(x_m, y, z, t), \tag{3.7}$$

which describes an instantaneous view of plume transport while having the same underlying form as commonly used models based on an ensemble-averaged view [65]. The key differences between this instantaneous view and common ensemble-averaged models are the time dependence of the distribution D that represents the stochastic nature of the turbulent plume, and the presence of $\delta_u(x_m, t)$ and $\delta_S(x_m, t)$ that accounts for non-stationarity in wind speed and mass storage in the control volume.

$D(x_m, y, z, t)$ is a random variable that captures the plume movement in time and the lateral and vertical directions as it responds to the instantaneous turbulent velocity components in these directions. It is therefore expected that $D(x_m, y, z, t)$ scales with the standard deviations of the velocity components in the y and z directions (σ_v and σ_w, respectively). It is well-understood that local scaling approaches based on MOST often provide an acceptable description of σ_w [45]. However, σ_v is affected by random large scale motions in the atmosphere that cannot be described accurately by local scaling laws [66]. Therefore, a greater degree of randomness is expected in y than in z directions for $D(x_m, y, z, t)$.

This observation highlights the benefits of integrating both sides of equation (3.7) with respect to y (i.e. across the plume), since the uncertainty associated with the lateral plume dispersion can be effectively removed [23]:

$$c^y(x_m, z, t) = \frac{Q_0}{u_e^M(x_m, t)} \left(\frac{\delta_S(x_m, t)}{\delta_u(x_m, t)} \right) D_z(x_m, z, t), \tag{3.8}$$

where $c^y(x_m, z, t) = \int_{y_{min}}^{y_{max}} c(x_m, y, z, t)dy$ is the cross-plume integrated concentration and $D_z = \int_{y_{min}}^{y_{max}} D(x_m, y, z, t)dy$ is a reflection of the vertical profile of mass concentration at x_m. Note that $D_z(x_m, z, t)$ is a random variable that is mainly driven by the stochastic nature of the vertical transport dynamics in the turbulent flow. In field applications, the sensor path is typically constrained by adjacent roadways, which are not always perpendicular to the wind direction. When the road segments are at a significant angle to the wind direction, $c^y(x_m, z, t)$ can be estimated by numerical integration of the mass concentration along the path as [23]:

$$c^y(x_m, z, t) = \sum_{y_{min}}^{y_{max}} c(x_m, y, z, t)\Delta t V \sin(\theta_r), \tag{3.9}$$

where Δt is the sensor acquisition time step, V is the vehicle velocity and $theta_r$ is the acute angle between the road segment and the wind direction.

We can account for all the stochasticity in $c^y(x_m, z, t)$ by introducing a fluctuating variable, $D_{z,e} = \frac{\delta_S(x_m,t)}{\delta_u(x_m,t)} D_z(x_m, z, t)$, because of the stochastic nature of $\delta_S(x_m, t), \delta_u(x_m, t)$ and $D_z(x_m, z, t)$. This new fluctuating variable is helpful in empirical analysis of the cross-plume integrated mass concentration and can be used to rewrite equation (3.8) as follows:

$$c^y(x_m, z, t) = \frac{Q_0}{u_e^M(x_m, t)} D_{z,e}(x_m, z, t). \tag{3.10}$$

Equation (3.10) can be used in a forward manner to estimate the downwind cross-plume integrated mass concentration $c^y(x_m, z, t)$ for a given emission rate

Q_0, as well as in the inverse problem of inferring Q_0 given downwind measurements of $c^y(x_m, z, t)$. Most dispersion models only offer an approximation of the ensemble-averaged $D_{z,e}(x_m, z, t)$, therefore, we apply Bayesian inference to account for the fluctuation of the instantaneous $D_{z,e}(x_m, z, t)$ from its ensemble averaged as detailed in section 3.2.2. For simplicity of notation the independent variables, x_m, z, and t will be dropped hereafter.

3.2.2 Bayesian inference for source estimation

Following Bayes' rule, and the notation introduced by Arumpalam et al. for recursive Bayesian inference [67], the posterior probability distribution of the emission rate Q based on the measurements of c^y at time step k (or after the k'th sensor pass) is [23], [61]

$$p\left(Q_k \middle| c^y_{1:k}\right) = \frac{p\left(Q_k \middle| c^y_{1:k-1}\right) p\left(c^y_k \middle| Q_k\right)}{p\left(c^y_{1:k}\right)}, \tag{3.11}$$

where $p\left(Q_k \middle| c^y_{1:k}\right)$, $p\left(Q_k \middle| c^y_{1:k-1}\right)$, and $p\left(c^y_{1:k}\right)$ are probability density functions (PDFs). $p\left(Q_k \middle| c^y_{1:k-1}\right)$ is the prior that is being updated through the recursion, $p\left(c^y_k \middle| Q_k\right)$ is the likelihood function, and $p\left(c^y_{1:k}\right)$ is the evidence term that ensures $p\left(Q_k \middle| c^y_{1:k}\right)$ integrates to unity. Note that the notation $c^y_{a:b}$ refers to the contiguous set of measurements between time (sensor pass) a and b inclusive.

In practical applications, past measurements of similar facilities [68] may be used to formulate the prior probability distribution at the first time step, i.e. before any sampling activities. Under the assumption that the only prior knowledge of Q is its lower and upper bounds, a uniform prior distribution can be adopted [61, 69]. This uniform distribution is often considered as sufficiently

uninformative based on the principle of maximum entropy [70] and can be expressed as follows

$$p(Q_1) = \frac{1}{Q_{max} - Q_{min}}, \tag{3.12}$$

where Q_{max} and Q_{min} are the prescribed upper and lower bounds of the emission rate, respectively.

The likelihood function, $p\left(c_k^y \middle| Q_k\right)$, describes the probability of observing c^y given Q at the k'th sensor pass and encodes all the information provided by the mass concentration measurements about the unknown emission rate [61]. Since the underlying distribution of the concentration measurements are unknown, the principle of maximum entropy supports the application of a Gaussian distribution with a prescribed error scale [71]. This choice for the likelihood function has proven useful in previous studies [23, 69, 72, 73], and is therefore adopted here. Furthermore, the dataset investigated in this study has been previously tested for leak estimation using the Gaussian likelihood function with satisfactory results [34] (more details on the dataset are provided in section 3.3) . Consequently, The Gaussian likelihood function in this study is expressed as

$$p\left(c_k^y \middle| Q_k\right) = \frac{1}{\sigma_e \sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{c_k^y - c_k^{y,M}(Q_k)}{\sigma_e}\right)^2\right], \tag{3.13}$$

where $c^{y,M}(Q) = \frac{Q}{u_e^M} D_z^M$ is the cross-plume integral of a modeled concentration for a given candidate value of Q. D_z^M is the estimated value of $D_{z,e}$ based on a Lagrangian Stochastic Model (LSM). The LSM is used to describe plume dispersion in a turbulent flow by modeling paths of fluid particles, which are driven by the random velocity field modeled by the generalized Langevin equation [74]. In this study, we impose the so-called well-mixed conditions and adopt Thomson's simplest solution for statistically stationary and horizontally homogeneous turbulence [75]. The LSM takes meteorological measurements (fric-

tion velocity, surface roughness, standard deviation of u and w and Obukhov Length) and the estimated distance between the emission source and the sensor as input parameters. The LSM is described in further detail in Appendix A.2.

In equation (3.13), σ_e is the uncertainty scale parameter, which can be estimated for observed data as

$$\sigma_e = \sqrt{\frac{1}{N-1}\sum_{k=1}^{N}\left(c_k^y - c^{y,M}(Q_0)\right)^2}, \tag{3.14}$$

where N is the number of passes per experiment. In this study, σ_e is estimated from the controlled release experiments and is known prior to the application of the Bayesian inference approach. In cases where σ_e cannot be estimated from prior measurements, it can be estimated using the error propagation method [25,76]. Briefly, σ_e is due to the following error scale parameters: 1) error due to the stochastic nature of atmospheric plume dispersion, 2) error due to the plume dispersion model and 3) measurements error including errors from the model input data. The parameterization of each of these error parameters is dependent upon the local meteorological conditions, the dispersion model used, and the quality of measurements.

The recursive Bayesian formulation of equation (3.11) used in conjunction with the uniform prior of equation does not have an analytical solution and should be solved numerically. For the numerical solution, Q is discretized from Q_{min} to Q_{max} with a uniform step size of ΔQ to form a vector of candidate values for Q to be considered. For each measurement of c^y, the likelihood function is evaluated at all candidate Q values using equation (3.13) and multiplied by the prior probability distribution. Subsequently, the evidence term after the k'th

sensor pass can be calculated through numerical integration as follows:

$$p\left(c_{1:k}^{y}\right) = \sum_{Q_{min}}^{Q_{max}} p\left(Q_k \middle| c_{1:k-1}^{y}\right) p\left(c_k^{y} \middle| Q_k\right) \Delta Q. \tag{3.15}$$

After calculating the evidence term, equation (3.11) can be applied to evaluate the posterior distribution of the emission rate at each candidate Q value. The same procedure is repeated after each mobile sensor pass and the posterior distribution is updated using a new prior (posterior at previous sensor pass) and a newly calculated likelihood function with the most recent c^y measurement.

After each sensor pass, the posterior PDF can be used to estimate the emission rate and the associated uncertainty. For instance, the emission rate after sensor pass k, can be calculated as the mean, median or mode of of the posterior PDF $p\left(Q_k \middle| c_{1:k}^{y}\right)$. Given that an uninformative, uniform prior distribution was adopted, we expect that the median and the mean to be heavily affected by the prior in the early stages of analysis. Therefore, the mode of the posterior PDF $p\left(Q_k \middle| c_{1:k}^{y}\right)$ is used as the estimated emission rate, i.e.:

$$E_k^Q = \arg\max_{Q_k} \left[p\left(Q_k \middle| c_{1:k}^{y}\right) \right]. \tag{3.16}$$

As number of mobile passes are increased, the effects of the prior distribution are reduced and the mode, median and mean of the posterior PDF grow closer in value. Furthermore, in cases where an informative prior can be derived prior from past experiments, the mean or median of the posterior PDF may be better candidates for the emission rate, as they better incorporate the prior information than the mode.

Finally, the associated uncertainty of the emission rate estimation using Bayesian inference is often calculated as the standard deviation, σ_k^Q, of the pos-

terior PDF:

$$\left(\sigma_k^Q\right)^2 = \int_{Q_{min}}^{Q_{max}} \left(Q - \overline{Q_k}\right)^2 \times p\left(Q_k \middle| c_{1:k}^y\right) dQ, \tag{3.17}$$

where Q_k is the expectation of the posterior PDF evaluated as follows:

$$\overline{Q_k} = \int_{Q_{min}}^{Q_{max}} Q \times p\left(Q_k \middle| c_{1:k}^y\right) dQ. \tag{3.18}$$

3.2.3 Bayesian inference for changepoint detection

In order to detect a change in the source emission rate, we apply the Bayesian Online Changepoint Detection (BOCD) methodology [62]. In this approach, changepoints are found by first estimating the posterior distribution over the run length, i.e. the time (or in this case, number of sensor passes) since the last changepoint, given the data observed so far. Denoting the length of the current run after sensor pass k with r_k, and applying the definition of conditional probability the run length posterior distribution can be expressed as

$$p\left(r_k \middle| c_{1:k}^y\right) = \frac{p\left(r_k, c_{1:k}^y\right)}{p(c_{1:k}^y)}, \tag{3.19}$$

where the run length evidence term is calculated using $p(c_{1:k}^y) = \sum_{r_k} p\left(r_k, c_{1:k}^y\right)$.

After every sensor pass there are two possibilities regarding the changepoint: 1) No changes occur after the sensor pass and therefore the run length is increase by 1 or 2) change occurs and run length is reset to 0. The probability that no changepoint occurs is referred to as the "growth probability" as it indicates that the run length is growing by 1 compared to the previous sensor pass. Similarly, we refer to the probability that a changepoint occurs as the "changepoint probability". We denote the growth probability such that the run length reaches $i \in \{0, 1, 2, ..., k-1\}$ after k sensor passes by $\alpha_k(i)$ and derive a recursive

estimate as follows

$$\begin{aligned}
\alpha_k(i) &= p\left(r_k = i, c^y_{1:k}\right) \\
&= p\left(r_k = i, r_{k-1} = i-1, c^y_k, c^y_{k-1}\right) \\
&= p\left(r_k = i, c^y_k \middle| r_{k-1} = i-1, c^y_{1:k-1}\right) p\left(r_{k-1} = i-1, c^y_{1:k-1}\right) \\
&= p\left(r_k = i, c^y_k \middle| r_{k-1} = i-1, c^y_{1:k-1}\right) \alpha_{k-1}(i-1). \qquad (3.20)
\end{aligned}$$

The first step in equation (3.20) is due to the fact that the run length can only increase by 1 after each sensor pass and the second step follows from the chain rule in probability theory. Furthermore, the recursive estimate in equation (3.20) requires the evaluation of $p\left(r_k = i, c^y_k \middle| r_{k-1} = i-1, c^y_{1:k-1}\right)$ which can be achieved by employing the chain rule again:

$$\begin{aligned}
p\left(r_k, c^y_k \middle| r_{k-1}, c^y_{1:k-1}\right) &= p\left(r_k \middle| r_{k-1}, c^y_{1:k-1}, c^y_k\right) p\left(c^y_k \middle| r_{k-1}, c^y_{1:k-1}\right) \\
&= p\left(r_k | r_{k-1}\right) p\left(c^y_k \middle| c^y_{k-i:k-1}\right), \qquad (3.21)
\end{aligned}$$

where $r_k = i$ and $r_{k-1} = i-1$ are implied and not explicitly written for simplicity of notation. The final step of equation (3.21) follows from the independence of the measured mass concentrations before and after a changepoint, which is true based on our assumption of independence of leak rates before and after a changepoint. Further, the condition on r_k in the second term on right hand side of the equation is absorbed by only limiting the conditional probability on measurements since the last changepoint leading to the subscript $k-i : k-1$. Equation (3.21) suggests that the growth probability can be computed based on two calculations: 1) The prior over r_k given r_{k-1} (also referred to as the changepoint prior) and 2) The predictive distribution over the new measurement after sensor pass k, given the data since the last changepoint.

The changepoint prior in equation (3.21) has nonzero mass at only two out-

comes, because after each sensor pass the run length either continues to grow such that $r_k = r_{k-1} + 1$ or a changepoint occurs and $r_k = 0$. Furthermore, the hazard function can be used to quantify each of these outcomes, since by definition the hazard function quantifies the probability that a changepoint occurs at a given time step conditioned that no changepoint has occurred prior to that time step [62,77]. Therefore, the changepoint prior can be expressed as

$$p(r_k|r_{k-1}) = \begin{cases} H(r_{k-1}+1) & \text{if } r_k = 0 \\ 1 - H(r_{k-1}+1) & \text{if } r_k = r_{k-1}+1 \\ 0 & \text{otherwise,} \end{cases} \tag{3.22}$$

where $H(r_{k-1})$ is the hazard function. The hazard function depends on the discrete a priori probability distribution over the interval between changepoints. However, in this study, we consider the special case where the a priori probability distribution is a discrete geometric distribution with timescale λ, i.e. the run length distribution is due to a memoryless process and the hazard function is constant at $1/\lambda$. The timescale λ can be set through prior knowledge, for instance, the average number of passes completed before a change in the emission rate occurs. In this study, λ is set to 15 based on the conducted experiments described in section 3.3.1.

The predictive distribution $p\left(c_k^y \middle| c_{k-i:k-1}^y\right)$ can be described by an equivalent distribution by utilizing the plume transport model of equation (3.10) leading to a one-to-one correspondence between $p\left(c_k^y \middle| c_{k-i:k-1}^y\right)$ and $p\left(Q_k \middle| c_{k-i:k-1}^y\right)$. In this case, the predictive distribution $p\left(c_k^y \middle| c_{k-i:k-1}^y\right)$ can be found through scaling of $p\left(Q_k \middle| c_{k-i:k-1}^y\right)$. We note that $p\left(Q_k \middle| c_{k-i:k-1}^y\right)$ is the prior distribution in equation (3.11), given the data since the last changepoint. Therefore, $p\left(Q_k \middle| c_{k-i:k-1}^y\right)$ and consequently $p\left(c_k^y \middle| c_{k-i:k-1}^y\right)$ can be estimated using the recursive Bayesian ap-

proach detailed in section 3.2.2, leading to simultaneous estimation of leak rate and changepoint detection. The computational details of calculating the growth probability term are provided in section 3.2.4.

The changepoint probability is evaluated in a similar manner to the growth probability. By noting that with the occurrence of a changepoint, the run length r_k drops to 0, we can derive a recursive estimate for the changepoint probability as

$$\begin{aligned}
\alpha_k(0) &= p\left(r_k = 0, c^y_{1:k}\right) \\
&= \sum_{j=1}^{k-2} p\left(r_k = 0, r_{k-1} = j, c^y_k, c^y_{k-1}\right) \\
&= \sum_{j=1}^{k-2} p\left(r_k = 0 | r_{k-1} = j\right) p\left(c^y_k \middle| c^y_{k-j-1:k-1}\right) \alpha_{k-1}(j),
\end{aligned} \tag{3.23}$$

where $\alpha_k(0)$ is the changepoint probability. The summation after the first step of equation (3.23) appears due to marginalization over the run length at sensor pass $k-1$. The intermediate steps in the derivation are omitted as they are identical to the derivation of the growth probability as outlined in equations (3.20) and (3.21). The estimation of each term on the right hand side of equation (3.23) follows the same procedure as equation (3.21).

Figure 3.2 illustrates the algorithm used to estimate the posterior distribution over the run length as described by equations (3.21) and (3.20). In this diagram, the solid blue lines correspond to growth probability calculations, while the red dashed lines are associated with changepoint probability evaluations. We note that multiple dashed lines arriving at a node in Figure 3.2, correspond to the marginalization over the run length at the previous sensor pass.

In practice, to automatically detect a changepoint and alert the system that

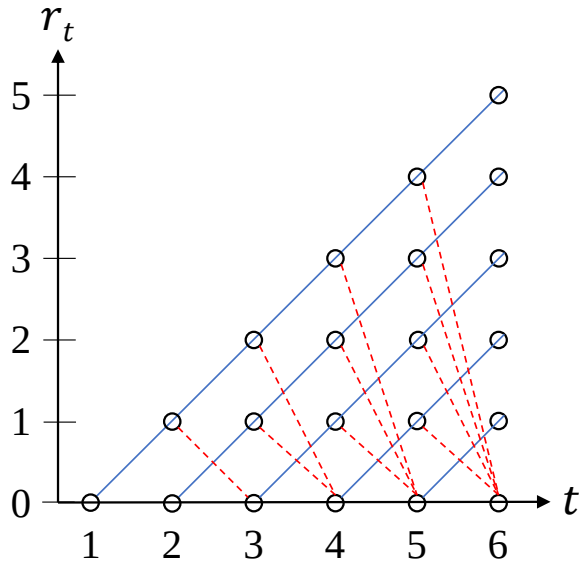

Figure 3.2: Visual description of the message passing algorithm used to estimate the run length distribution over the observed data. The circles represent run length hypotheses and the lines between the circles show recursive transfer of mass between sensor pass (or time step). Solid lines indicate that probability mass is being passed upwards, causing the run length to grow at after the next sensor pass and dashed lines indicate that the current run is truncated, and the run length drops to zero.

a change in leak rate has occurred a prescribed condition on the changepoint probability should be put into place. The appropriate detection condition can be chosen based on the application and the available ancillary information regarding the emission conditions. In this study, we use a changepoint probability threshold such that when the changepoint probability is above this threshold the system automatically registers a change in leak rate, the posterior probability over the emission rate at the previous sensor pass is retained and the prior distribution in equation (3.11) is reset to the uniform prior of equation (3.12).

3.2.4 Computational details of growth probability estimation

The calculation of the growth probability distribution after sensor pass k relies on knowledge of the probability distribution of the emission rate conditioned on the observed data from the last changepoint until the previous sensor pass at $k-1$, i.e., $p\left(Q_k \middle| c^y_{k-i:k-1}\right)$ as shown by the derivation in equation (3.21). Furthermore, the growth probability is calculated for every value of run length $i \in \{0, 1, 2, ..., k-1\}$, which requires the recursive calculation of the posterior probability of emission rate for every i through equation (3.11). Therefore, the growth probability calculation can be computationally expensive for large values of k. To overcome this challenge, we note that the likelihood function in equation (3.11) is independent of i, hence, the calculation of the posterior probability of emission rate for every i should completed in a vectorized or Single Instruction, Multiple Data (SIMD) manner that allows for parallel processing of data and significantly improves the run time of the probability estimation [78].

3.3 Materials and Methods

In this section we first describe the conducted field experiments that yielded the data used to examine the Bayesian framework. Throughout these experiments, the leak rate was kept constant, therefore, a data synthesis procedure is implemented to simulate a step change in the leak rate before the application of the changepoint detection algorithm. A series of performance measures are then defined to quantify the performance of the changepoint detection algorithm.

3.3.1 Field experiments

Experiments of controlled releases of methane were conducted at the McGovern soccer training field of Cornell University (Game Farm Rd, Ithaca, NY. USA) in early August 2016. During the experiments, the site was covered with short grass (~5cm) and located in a relatively open field with a ~500m distance from a residential area in the West, ~150m distance from small forest in the North and approximately 400m (500m) distance from roads on the east (south) side. Point-source emission of methane (99.9% pure gas) was controlled by a mass flow controller (SmarTrak 100 from Sierra Instruments Inc., Monterey, CA, USA), with a mass flow accuracy of ±1%. The height of the methane release was similar to the height of the grass at 5cm. On the west side of the field, a 3D sonic anemometer (CSAT-3, Campbell Scientific Inc., Logan, UT, USA) was installed on a small tower to measure local meteorological conditions. The height of the tower was 2.31m with the sonic anemometer measuring the three components of wind velocity and air temperature at a frequency of 10 Hz. To mimic emissions from a source surrounded by other low-level structures, a 1.4m barrier (windbreak) was established in a circle around the emission source. This setup can for instance, approximate a well head located in densely organized well pad or a pipeline within a natural gas metering station.

A mobile measurement platform (MMP) was configured with a precise GPS unit (Trimble Geo 7X handheld from Trimble Inc., Sunnyvale, CA, USA) to track its position at a sampling frequency of 1 Hz. The accuracy of the GPS unit was approximately 5-15 cm for >97% of the measured data points after post processing. The MMP was equipped with a LI-COR LI-7700 open-path methane analyzer (LI-COR Biosciences, Lincoln, NE, USA), that outputs methane mixing

ratios in the unit of parts per million (ppm). The operating frequency of the analyzer was set to 10 Hz, and it was positioned at a height of 1.3m. Furthermore, the analyzer was calibrated by the manufacturer less than a month before the experiment, and is designed with an open-path configuration for long term monitoring without regular re-calibration.

Regarding the measurements, a conversion factor is applied to translate above-ambient mixing ratios (c_a, in ppm) into mass concentrations (c, in g/m^3) in equation (3.2). This conversion factor is dependent on the molecular weight of the released gas (16.04 g/mol for methane) and the ambient temperature which affects the molar of the gas. The above-ambient mixing ratios are found by subtracting the ambient methane mixing ratios from the raw methane mixing ratios measured by the open-path analyzer, with the ambient mixing ratio calculated as the 5th percentile of the ranked time series of raw mixing ratio measurements [23,34,68,79]. The estimated ambient mixing ratio was compared to methane mixing ratios measured prior to the experiments with minimal differences found (<2%), suggesting that the ambient mixing ratio was determined robustly.

To ensure perpendicular sensor passes with respect to the wind direction, stake flags were placed in three circles centered at the emission source with radii of 10, 20 and 30m, and repeated passes were made along each of the circles. In addition, the start of a pass took place approximately one minute after the end of the previous pass, warranting the independence of the measurements of each pass from those of previous passes. The average sensor speed was very low during the experiments (approximately 2 m/s) to better capture the plume structure. Data was aggregated within 30 minute intervals, during which the

measurements from the meteorological tower was used to estimate meteorological parameters such as the Obukhov length (L) and the friction velocity (u_*). Each set of passes completed in a 30-minute period is considered and analyzed as a single experiment.

3.3.2 Data synthesis

A total of 18 experiments were conducted with the 1.4m barrier present in the field, with six experiments at each source-to-sensor distance (i.e. x_m of 10, 20 and 30m). Two selection requirements were established to filter out experiments performed under unacceptable conditions. First, experiments under stable atmospheric conditions were excluded, and only experiments conducted under neutral or unstable conditions were retained. Second, experiments conducted under low wind ($\bar{u}$ < 1.0 m/s) and high turbulent intensity (I_u >0.5) conditions were discarded. As a result of this selection criteria, 14 experiments were retained for further analysis, the details of which are summarized in Table 3.1 alongside the meteorological conditions reported by the meteorological tower.

In all the conducted experiments, the emission rate was kept constant at a rate of Q_0 = 0.083 g/s. Therefore, a data synthesis procedure was established to artificially simulate a change in the emission rate. The proposed data synthesis approach relies on the observation that the order of the measurements of the cross-plume integrated mass concentrations through each pass of the MMP can affect the performance of the changepoint detection algorithm. For instance, consider two permutations of the measurements of the same experiment (experiment ID 14 in Table 3.1) as shown in Figure 3.3. In the first permutation, series

Table 3.1: Summary of experimental conditions, including experiment identification number (ID), approximate source-to-sensor distance (x_m), number of sensor passes for the experiment (N), and sampling day of year (DOY), and meteorological conditions as measured by the nearby meteorological tower, including mean streamwise velocity ($\bar{u}$), standard deviation of streamwise velocity (σ_u), turbulent intensity (I_u), friction velocity (u_*), mean wind direction (θ_m) clockwise from north, sensible heat flux (H), and atmospheric stability (z/L). The meteorological variables are derived from data collected during the corresponding experiments (30 min).

ID	x_m **(m)**	N	**DOY**	$\bar{u}$ **(m/s)**	σ_u **(m/s)**	I_u **(-)**	u_* **(m/s)**	θ_m **(deg)**	H **(W/m^2)**	z/L **(-)**
1	30	14	217	2.94	0.98	0.28	0.18	147	77.67	-0.32
2	20	16	217	2.72	1.06	0.31	0.23	149	148.97	-0.31
3	10	15	217	2.49	0.98	0.33	0.29	172	138.13	-0.14
4	30	12	217	2.72	1.15	0.31	0.24	152	161.21	-0.30
5	20	16	217	2.95	1.08	0.29	0.22	146	171.01	-0.41
6	10	16	217	2.48	0.97	0.33	0.21	159	159.46	-0.46
7	30	14	218	3.41	1.34	0.28	0.37	204	219.77	-0.11
8	20	14	218	3.70	1.27	0.26	0.36	211	223.92	-0.12
9	10	13	218	3.82	1.33	0.26	0.38	212	207.86	-0.10
10	30	16	218	4.18	1.31	0.24	0.36	194	171.36	-0.09
11	20	13	218	4.31	1.31	0.24	0.37	184	173.08	-0.08
12	10	13	218	3.99	1.23	0.25	0.40	179	129.53	-0.05
13	20	12	219	2.75	1.04	0.30	0.16	318	125.83	-0.77
14	10	13	219	2.29	1.04	0.35	0.20	315	140.29	-0.47

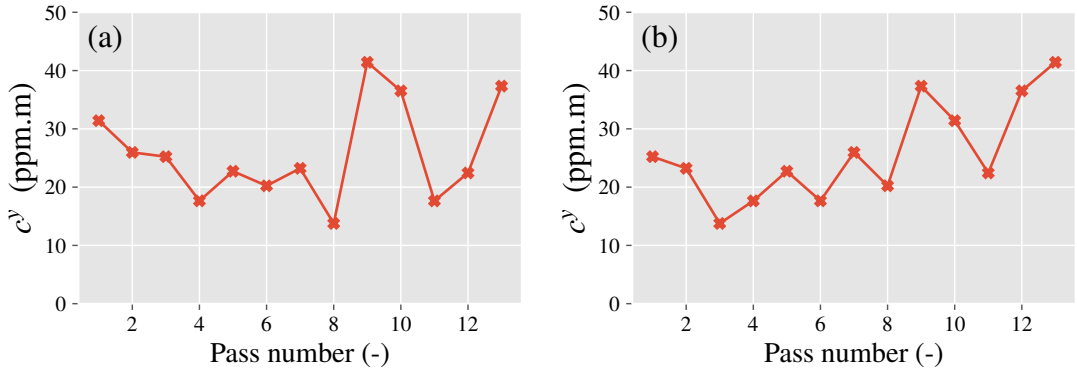

Figure 3.3: Two instances (a) and (b) of the same experiment (experiment ID 14) created through random shuffling of the measurements. The measurements are cross-plume integrated above-ambient mass concentrations of methane calculated over sensor passes.

of measurements leading to multiple low c^y values are followed by a comparatively high measurement (at sensor pass 9), therefore the algorithm is likely to detect a changepoint while in reality the emission rate has been constant. However, in the second permutation, the difference in consecutive c^y measurements are smaller than the first permutation and the probability of detecting a changepoint is lowered.

This observation motivates the use of random shuffling in synthesizing experiments consisting of a change in the leak rate after a few passes of the MMP around the point source. Given the independence of each c^y measurement (as discussed in section 3.3.1), random shuffling does not lead to any loss of information (e.g., information regarding correlation between measurements). For each experiment, the data synthesis steps are as follows (Experiment ID 4 in Table 3.1 used in Figure 3.4):

1. For each pass, c^y is calculated to create a time series for the experiment, which will be referred to as the "original signal" as depicted in Figure

3.4a.

2. The original signal is duplicated and scaled by a given constant that is chosen based on the ratio of the leak rate before and after a simulated changepoint resulting in a ”scaled signal” as presented in Figure 3.4b.
3. Both original and scaled signals are shuffled to create a random permutation of each respective signal with the results shown in Figure 3.4c and 3.4d
4. The two shuffled signals are concatenated such that the first measurement of the shuffled scaled signal follows the last measurement of the shuffled original signal, creating a signal that consists of a changepoint as illustrated in Figure 3.4e.

Steps 2-4 of the above procedure are repeated 1000 times, to create a total of 1000 signals with changepoints for each experiment. The changepoint detection algorithm is then applied to these signals and the performance of the algorithm is evaluated according to the performance measures described in section 3.3.3.

3.3.3 Performance measures

The performance of changepoint detection methods are often evaluated through a series of commonly used measures. The importance of each measure is dependent upon the application of the changepoint detection system. Here, we introduce four different performance measures, that are slightly altered with respect to common definitions to better suit the context of changepoint detection in emission rates. In the following description of the performance measures,

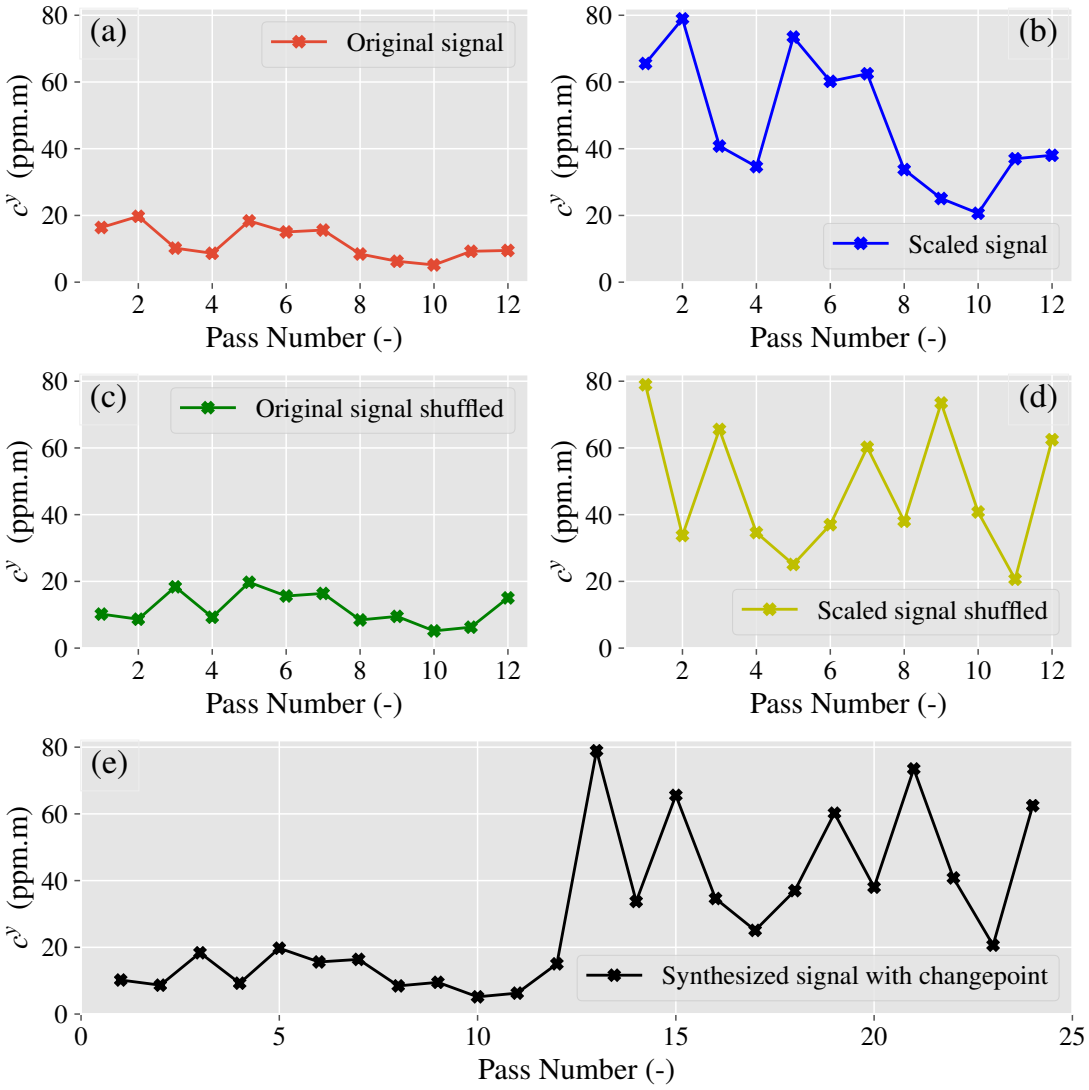

Figure 3.4: Summary of the data synthesis procedure for one instance of an experiment (ID 4), where the (a) original measurements from the field experiment are (b) scaled by multiplying all measurements by a prescribed constant. Two sets of new measurements are created by random shuffling leading to a (c) shuffling of the original measurements and a (d) shuffling of the scaled measurements. The two shuffled sets are then (e) concatenated to create one instance of synthesized measurements consisting of a step change.

each unique signal consisting of a changepoint is referred to as an ”instance” of an experiment.

Recall This refers to the fraction of the changepoints that are detected after exactly one sensor pass following the change in emission rate. We label these successful changepoint detections as ”True Positive” instances denoted by TP. Instances where the changepoints are detected with a delay, i.e. where changepoints are detected after at least two sensor passes after the change in emission rate are labeled ”Delayed True Positive” (DTP), and instances where changepoints were not detected are referred to as ”False Negative” (FN) instances. Therefore, recall which is a measure of how effective the changepoint algorithm is in detecting changepoints as soon as they occur is expressed as

$$Recall = \frac{TP}{TP + DTP + FN}. \tag{3.24}$$

Detection Recall This refers to the fraction of changepoints that are detected any time after the change in emission rate has occurred. Employing the labels introduced earlier, detection recall as a measure of how effective the changepoint algorithm is in detecting the changepoints is expressed as

$$Detection\ Recall = \frac{TP + DTP}{TP + DTP + FN}. \tag{3.25}$$

Detection Delay This measures the average number of passes that it takes to detect the changepoint after the emission rate has changed. This measure is evaluated only for experiments where the changepoints were detected for all instances (with or without delay) and is evaluated as

$$Detection\ Delay = \frac{\sum_{i=1}^{TP+DTP} Predicted\ CP - Actual\ CP}{TP + DTP}, \tag{3.26}$$

where *Predicted CP* refers to the sensor pass after which the changepoint is detected and *Actual CP* refers to the sensor pass after which the change in leak rate has occurred.

False Positive Rate This refers to the ratio of number of instances where changepoints are detected prior to the change in emission rate to to total number of instances. Here "False Positive" (FP) refers to instances where data points that are not changepoints are recognized as changepoints. The False Positive Rate is a measure that reflects how many false alarms would be generated by the changepoint detection algorithm and is expressed as follows

$$False\ Positive\ Rate = \frac{FP}{All\ instances} = \frac{FP}{1000}, \tag{3.27}$$

noting that *All instances* refers to the 1000 signals with changepoints created for each experiment.

To account for the variability introduced through random shuffling during the data synthesis stage, for each experiment the data synthesis procedure is repeated 100 times and therefore 100 different estimates of each performance measure are computed. These 100 values of the performance measures are then collected and used in a bootstrapping significance test analysis to establish 95% confidence intervals for the computed performance measures [80].

3.4 Results and Discussion

Before presenting the results related to the performance of the changepoint detection algorithms across all experiments, we explore one instance of an experiment. For this instance, it is shown how the changepoint detection algorithm is

coupled with the emission estimation procedure to approximate the leak rates before and after a change in the emission rate.

3.4.1 Leak estimation and changepoint detection

We first show the changepoint detection procedure for one instance of an experiment (Experiment ID 4 in Table 3.1). In this instance, 12 sensor passes are made before the leak rate is significantly increased from Q_1 = 0.083 g/s to Q_2 = 0.332 g/s which is four times as large as Q_1. The c^y measurements including the changepoint are shown in Figure 3.5a. After each sensor pass, the changepoint probability is evaluated through the procedure described in section 3.2.3 with the values presented in Figure 3.5b. A changepoint is detected when the changepoint probability surpasses a prescribed probability threshold of 0.8, after which the prior to the recursive Bayesian inference of equation (3.11) is reset to the uniform prior (equation(3.12)), so that the new emission rate can be approximated. Furthermore, the figure shows an increase in the changepoint probability after the 19th sensor pass which can be attributed to the high value of the c^y measurement (in comparison to previous measurements) corresponding to this sensor pass. In this study, the changepoint probability threshold is chosen through trial and error to lower the false alarm rate of the detection algorithm, and its effect on false positive rate is investigated in section 3.4.2.

For the same experiment instance as above, Figure 3.6 illustrates the evolution of the posterior PDF of the emission rate after each sensor pass before and after the changepoint. In this case, the lower and upper bounds of the emission rate, denoted by Q_{min} and Q_{max} are specified as 0 and 5.0 g/s. The choice for

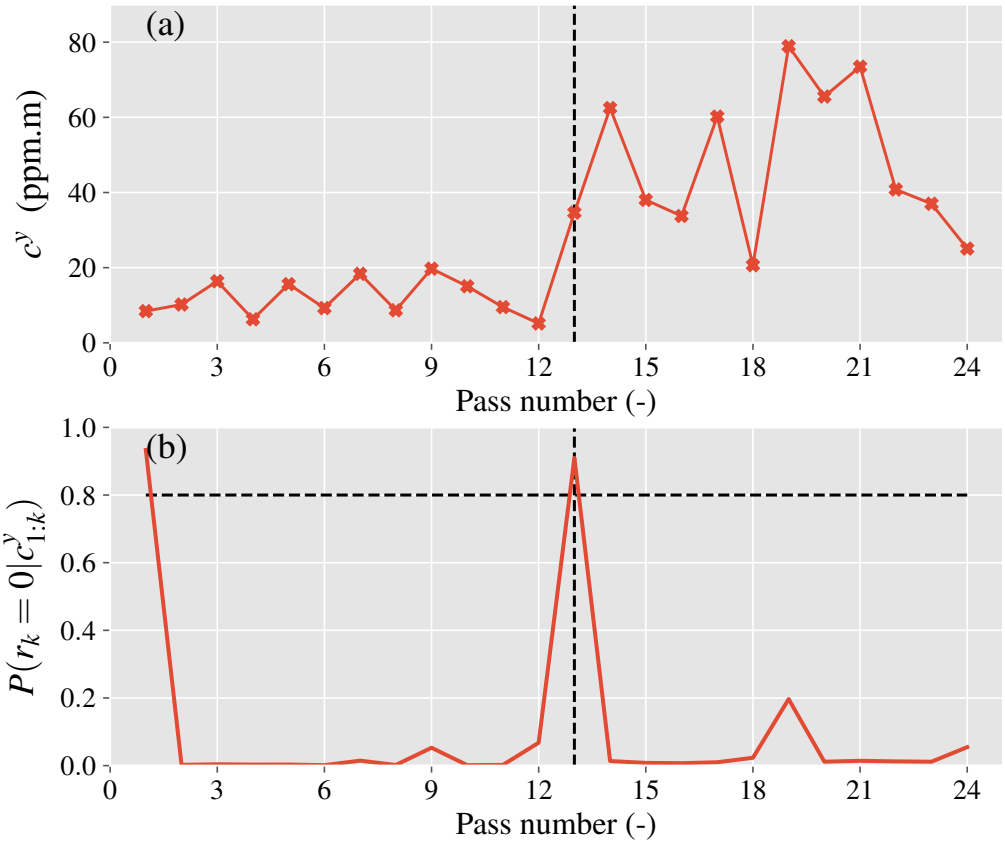

Figure 3.5: Application of the changepoint algorithm to an instance of an experiment (ID 4). (a) Synthesized instance with a step change in leak rate from 0.083 g/s to 0.332 g/s, where the vertical dashed line indicates the first sensor pass after the change. (b) The changepoint probability plotted after every sensor pass, where the horizontal dashed line represents the changepoint threshold probability, above which the algorithm registers a changepoint and resets the recursive Bayesian inference for leak estimation.

Q_{min} is trivial as the emission rate can only take positive values. Q_{max} is determined through trial and error such that the tail of the derived posterior PDF of the emission rate is close to zero. Using a larger Q_{max} does not affect the accuracy of the Bayesian inference procedure, however it is deemed unnecessary as it increases the computational cost of the recursive Bayesian inference scheme. Figure 3.6b shows that posterior PDF is fairly small at Q = 2.0 g/s, suggesting

that the choice of Q_{max} = 5.0 g/s is effective. Starting from a relatively broad posterior PDF, suggesting a large uncertainty in the emission rate, the posterior PDF tends to approach a more narrow shape with additional sensor passes. It is worth noting that after the change in emission rate, the variation in the c^y measurements are much larger compared to measurements at the original emission rate. Therefore, it is necessary to use a new estimate for σ_e in equation (3.13) for approximating the emission rate after the changepoint. In practice, the emission rate after the change is not known, hence, a larger and more conservative choice for σ_e can be used to accommodate this lack of information. In the example shown in Figure 3.6, we employ $\sigma_{e,2} = 10 \times \sigma_{e,1}$ where $\sigma_{e,1}$ and $\sigma_{e,2}$ refer to the error scale parameters before and after the changepoint, respectively, which is a conservative choice given that the error scale after the change is four times the error scale prior to the change. This conservative choice leads to a higher projected uncertainty when it comes to estimating the emission rate after the changepoint.

3.4.2 Changepoint detection performance

We investigate the performance of the changepoint detection method using the measures introduced in section 3.3.3 by systematically varying the magnitude of the change in leak rate when synthesizing the data. To this end, Figure 3.7 shows recall for varying values of "jump-to-noise ratio" (JNR), where JNR is the ratio of the absolute difference in the average c^y before and after the change (i.e., the jump) to the standard deviation of c^y before the change in leak rate (i.e., the noise). As expected, when the change in leak rate is of the order of the noise in the measurements, or in other words JNR is of the order of 1, changepoints

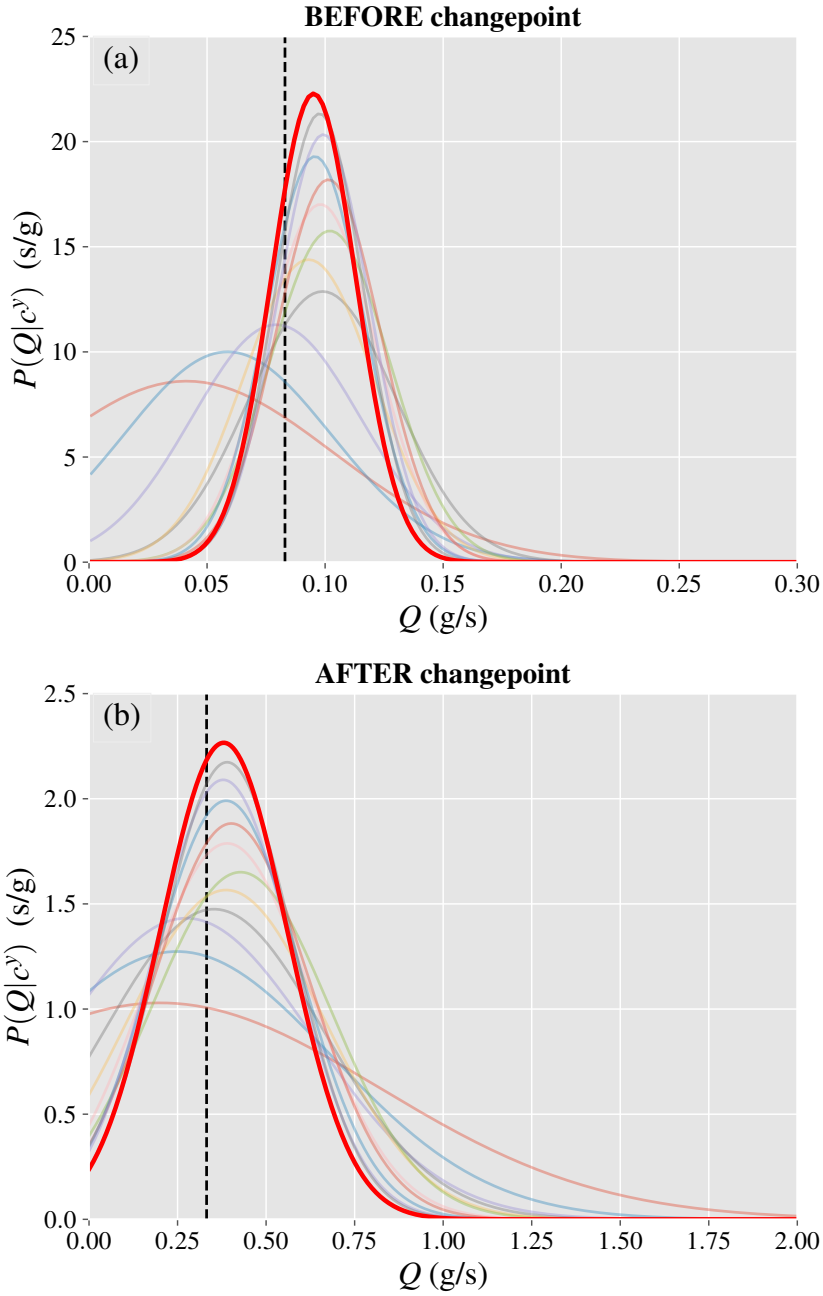

Figure 3.6: The evolution of the posterior probability $p(Q|c^y)$, of the emission rate Q after each sensor pass (a) before and (b) after the change in leak rate as detected by the changepoint detection algorithm for one instance of an experiment (ID 4). The posterior probability obtained after the final sensor pass before the changepoint in (a) and after the overall final pass are presented with a solid red line. The vertical dashed lines indicate the actual emission rate of (a) 0.083 g/s and (b) 0.332 g/s.

are difficult to detect and therefore recall is low for all experiments. Further, as JNR is increased a monotonic rise in performance is observed across all experiments, with similar recall values observed in almost all cases. This similarity of recall values across experiments for each JNR motivates the idea of grouping all experiments based on the source-to-sensor distance. Figure 3.8 presents the recall averaged across all experiments within each group as a function of JNR, where the vertical bars indicate the 95% confidence intervals. The trends observed in Figures 3.7 and 3.8 suggest that JNR can solely predict the recall for the changepoint algorithm irrespective of the source-to-sensor distance and the measurement noise.

In practice, it is more constructive to predict the performance of the changepoint algorithm based on the ratio of the emission rate before and after the changepoint. To this end, Figure 3.9 illustrates recall as a function of increasing leak rate ratio (LRR) for all experiments, where leak rate ratio is the ratio of the leak rate after the change to the leak rate before the change. In this figure, for each source-to-sensor distance, the experiments are sorted based on coefficient of variation (CV) of c^y. For each experiment, CV is calculated as the ratio of the standard deviation of c^y measurements to the average c^y measurements in the original signal (e.g., Figure 3.4a). It can be seen that a higher CV is a predictor for lower recall as an indicator for the performance of the changepoint algorithm. This relationship between recall and CV can be explained through a comparison between LRR and JNR.

According to equation (3.10), c^y is directly proportional to the leak rate Q, therefore in our synthesized data, the leak rate ratio is the same as the ratio of

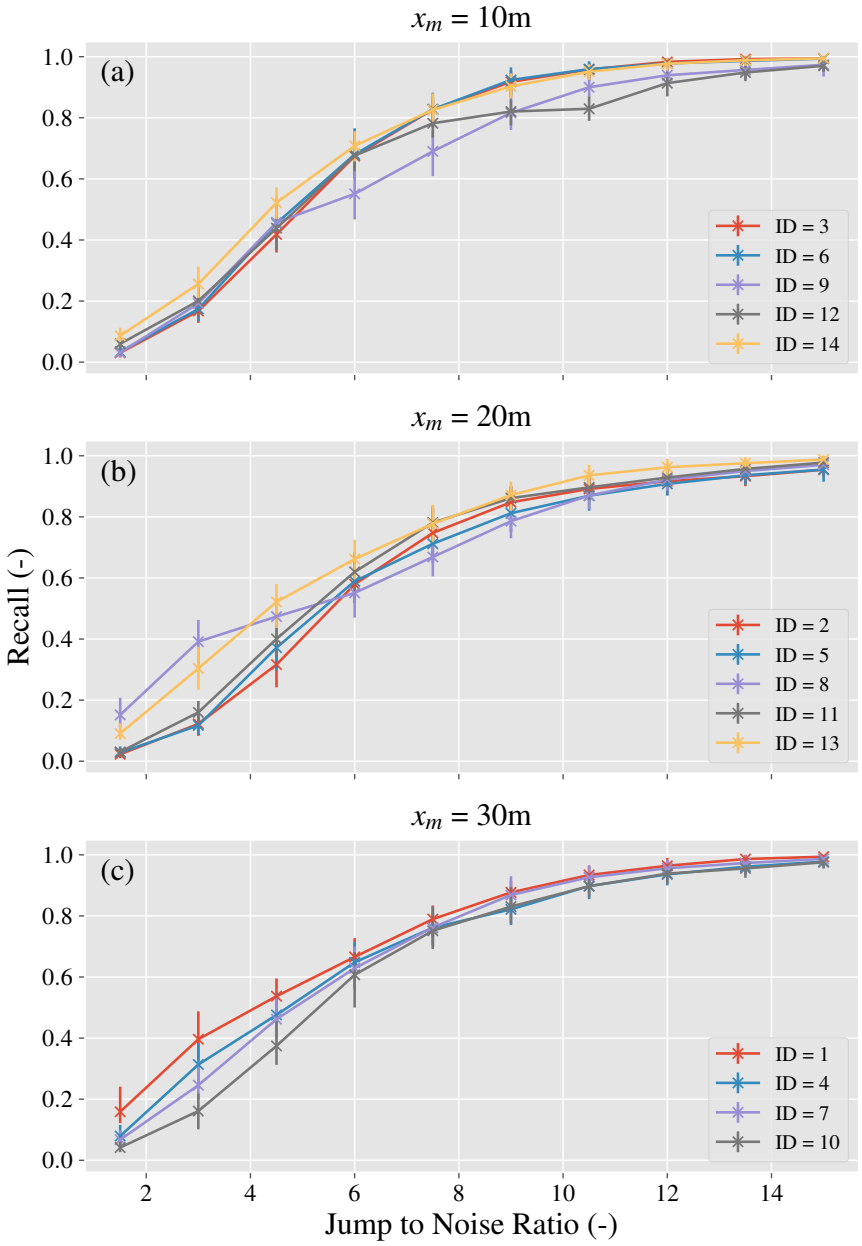

Figure 3.7: Evolution of recall for a series of jump to noise ratios varying between 1.5 and 15.5 for source-to-sensor distances, x_m of (a) 10m, (b) 20m and (c) 30m. ID refers to the experiment ID as seen in Table 3.1. Vertical bars represent the 95% confidence intervals.

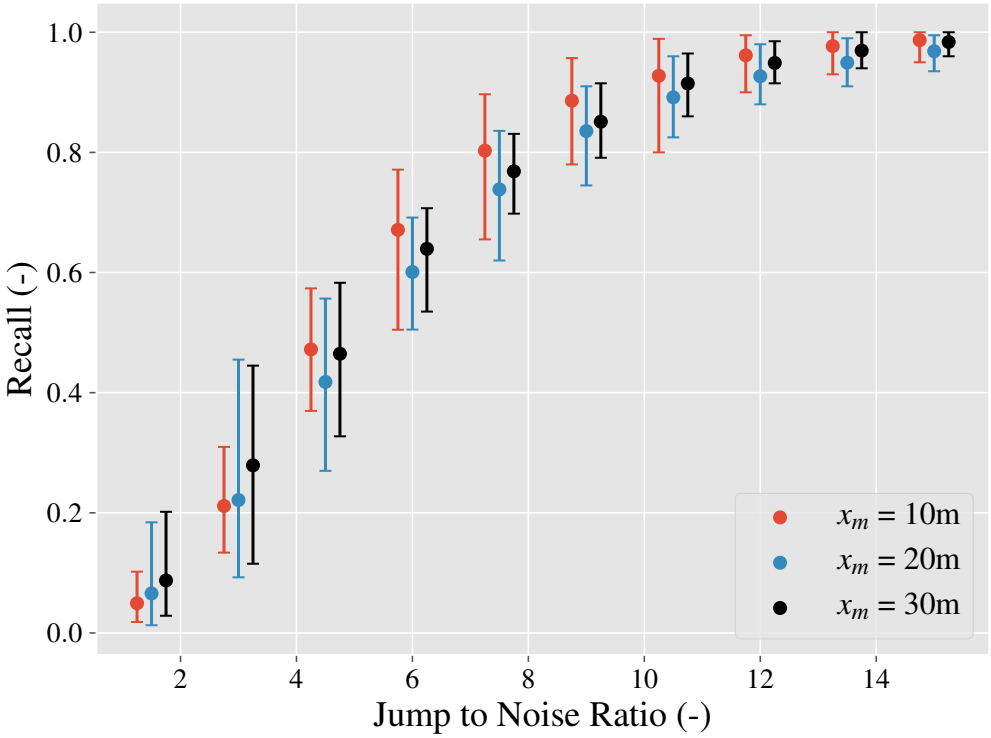

Figure 3.8: The evolution of recall for a series of jump to noise ratios varying between 1.5 and 15.5 after grouping experiments based on source-to-sensor distance, x_m. For each x_m, the set of jump to noise ratios are identical, however they are plotted in an offset to improve visibility. Vertical bars represent the 95% confidence intervals.

the mean c^y after and before the change. Therefore we can write

$$LRR = \frac{Q_2}{Q_1} = \frac{\mu_{c_2}}{\mu_{c_1}}, \tag{3.28}$$

where Q_2 and Q_1 are the emission rate after and before the change, and μ_{c_2} and μ_{c_1} are the average c^y measurements after and before the change, respectively. With this definition, we can relate JNR and LRR as follows

$$JNR = \frac{\mu_{c_2} - \mu_{c_1}}{\sigma_c} = \frac{LRR - 1}{CV}, \tag{3.29}$$

where σ_c is the standard deviation of c^y measurements in the original signal of an experiment. Based on equation (3.29), for a constant LRR, a higher value

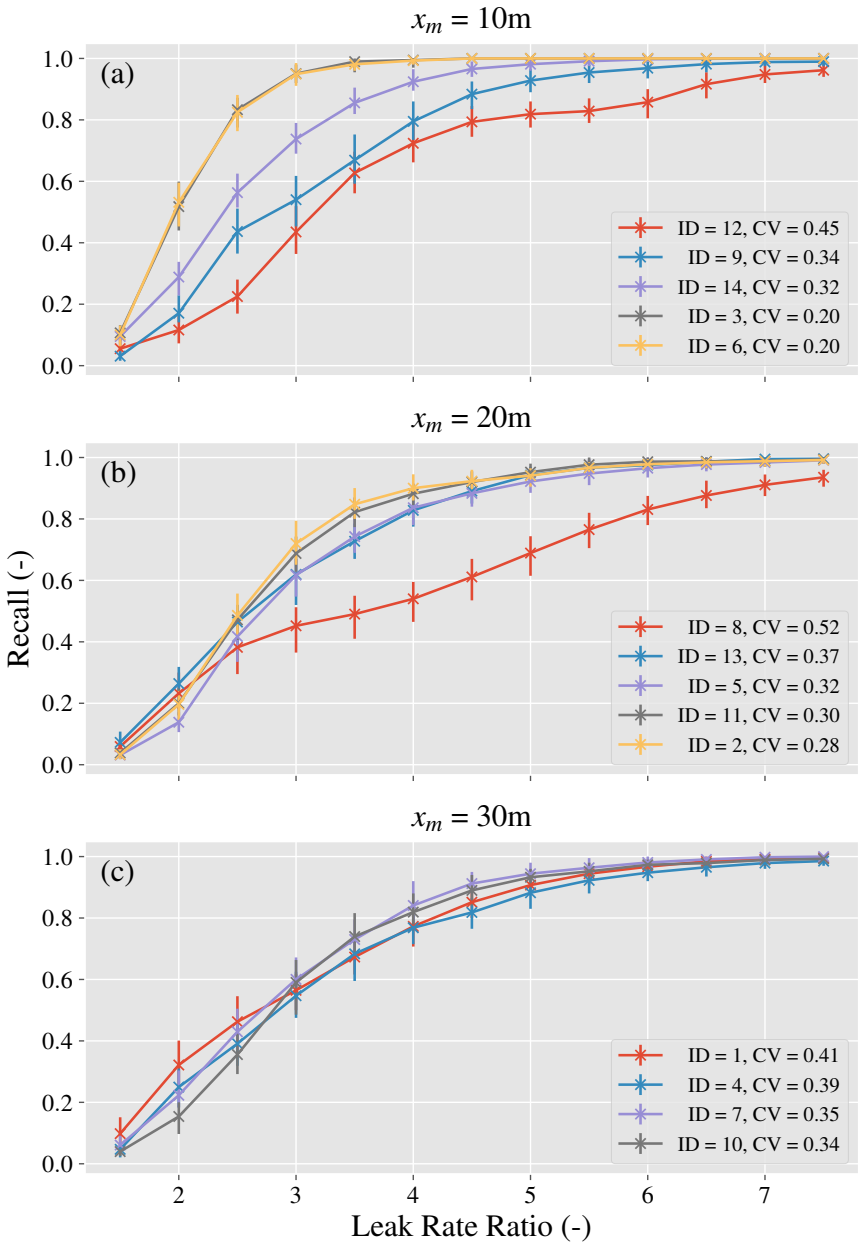

Figure 3.9: Evolution of recall for a series of leak rate ratios varying between 1.5 and 7.5 for source-to-sensor distances, x_m of (a) 10m, (b) 20m and (c) 30m. ID refers to the experiment ID as seen in Table 3.1, and CV refers to coefficient of variation of c^y measurements calculated for each experiment. Vertical bars represent the 95% confidence intervals.

of CV corresponds to a smaller JNR, which according to Figure 3.8 points to a lower recall.

In most practical applications, delayed detection of the changepoint is acceptable. Therefore, Figure 3.10 depicts the detection recall as a function of increasing LRR. In this figure, for each source-to-sensor distance, the experiments are sorted based on σ_c as a measure of noise in the measurements. It can be seen that experiments with higher values of σ_c correspond to higher detection recalls. This behaviour is expected due to our data synthesis procedure, where high values of σ_c lead to significantly large c^y measurements after the change in emission rate which are easily detected by the changepoint detection algorithm. Moreover, above a leak rate ratio of 3, the changepoints are rarely missed if we account for delayed detection, therefore showing the effectiveness of the algorithm in raising the alarm when a substantial change in the emission rate occurs. The significance of this result can be highlighted by noting that in a recent study of natural gas well pads in California, it was shown that well pads for which facility-based emission estimates were at least 3 times the component-based emission estimations, were responsible for 80% of the total measured emissions from all well pads [2].

Given that in some applications the change in emission rate can be intermittent, it is also important to quantify the delay in changepoint detection. Therefore, Figure 3.11 illustrates the detection delay for experiments where the changepoints are successfully detected across all instances (i.e., Detection Recall = 1) against increasing leak rate ratio. In this case, there is no clear trend between the noise in the measurements and the detection delay. However, the delay in changepoint detection monotonically decreases with increasing LRR as

expected. It is worth noting that even at the lowest LRR where all changepoints are detected, the detection delay is less than one sensor pass, showcasing the speed of the changepoint detection algorithm.

Finally, we investigate the sensitivity of the changepoint detection algorithm to the changepoint probability threshold, which in earlier results was set to a value of 0.8. Figure 3.12 presents the false positive rate when varying the changepoint probability threshold from 0.5 to 0.95. It can be seen that even at the lowest chosen threshold the false positive rate is less than 12%, highlighting the robustness of the changepoint detection algorithm. Furthermore, while the probability of false alarms is generally higher for experiments with larger noise, noise is not the sole predictor of the false positive rate in experiments. Range of the measurement distribution in each experiment, i.e., the difference between the maximum and minimum c^y measurements in each experiment, seems to be a better predictor than standard deviation of measurements for the false positive rate. Consequently, the changepoint detection algorithm can be adversely affected by the presence of outliers in the data. There are multiple possible solutions for alleviating the sensitivity of the changepoint detection algorithm to outliers. One possible solution is to modify the changepoint detection condition first introduced in section 3.2.3. For example, the condition can be adapted such that a changepoint is retained only if the changepoint probability is above a threshold for multiple measurements over the next few sensor passes. This requires the algorithm to delay resetting the Bayesian inference of equation (3.11) until the detection condition is satisfied. The downside of this solution is potential poor changepoint detection when the change in emission rate is intermittent and temporary.

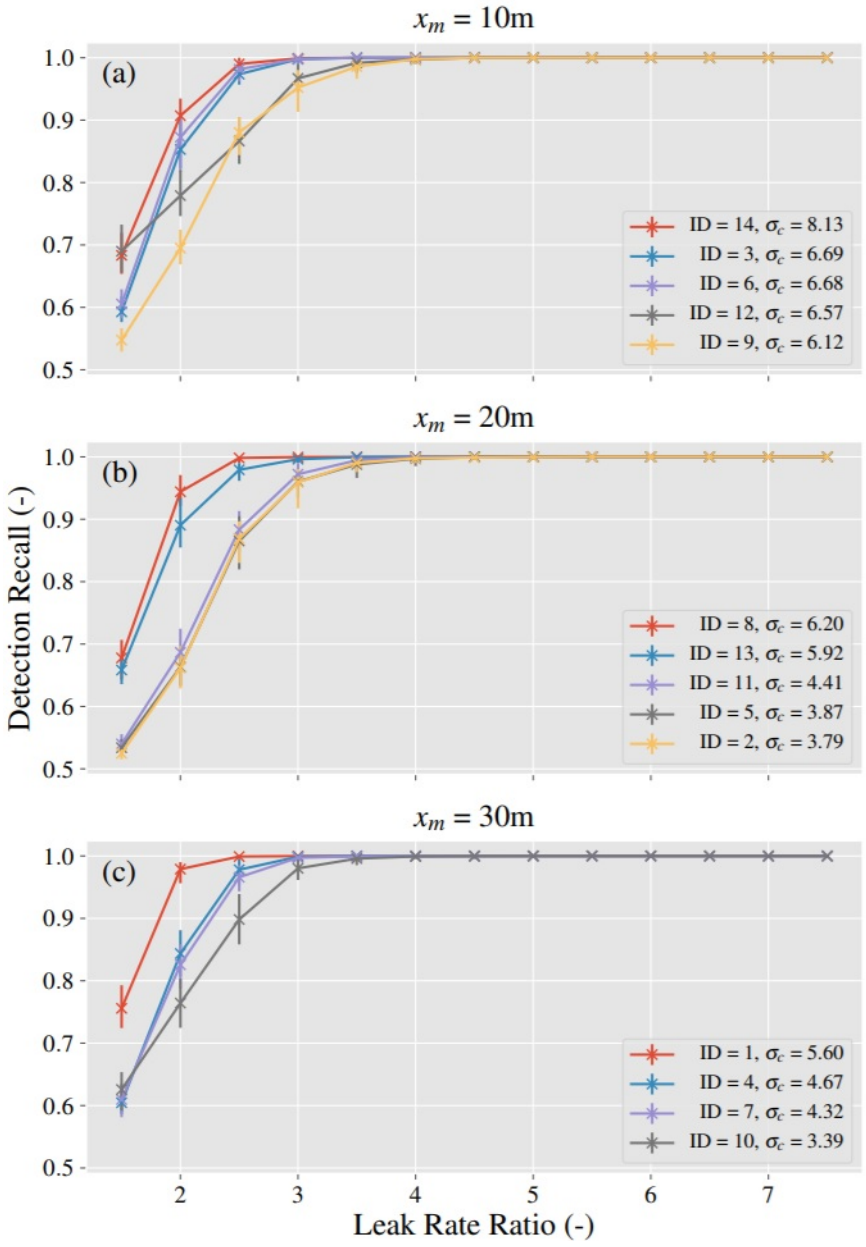

Figure 3.10: Evolution of detection recall for a series of leak rate ratios varying between 1.5 and 7.5 for source-to-sensor distances, x_m of (a) 10m, (b) 20m and (c) 30m. ID refers to the experiment ID as seen in Table 3.1, and σ_c refers to the standard deviation of c^y measurements before the changepoint that is calculated for each experiment. Vertical bars represent the 95% confidence intervals.

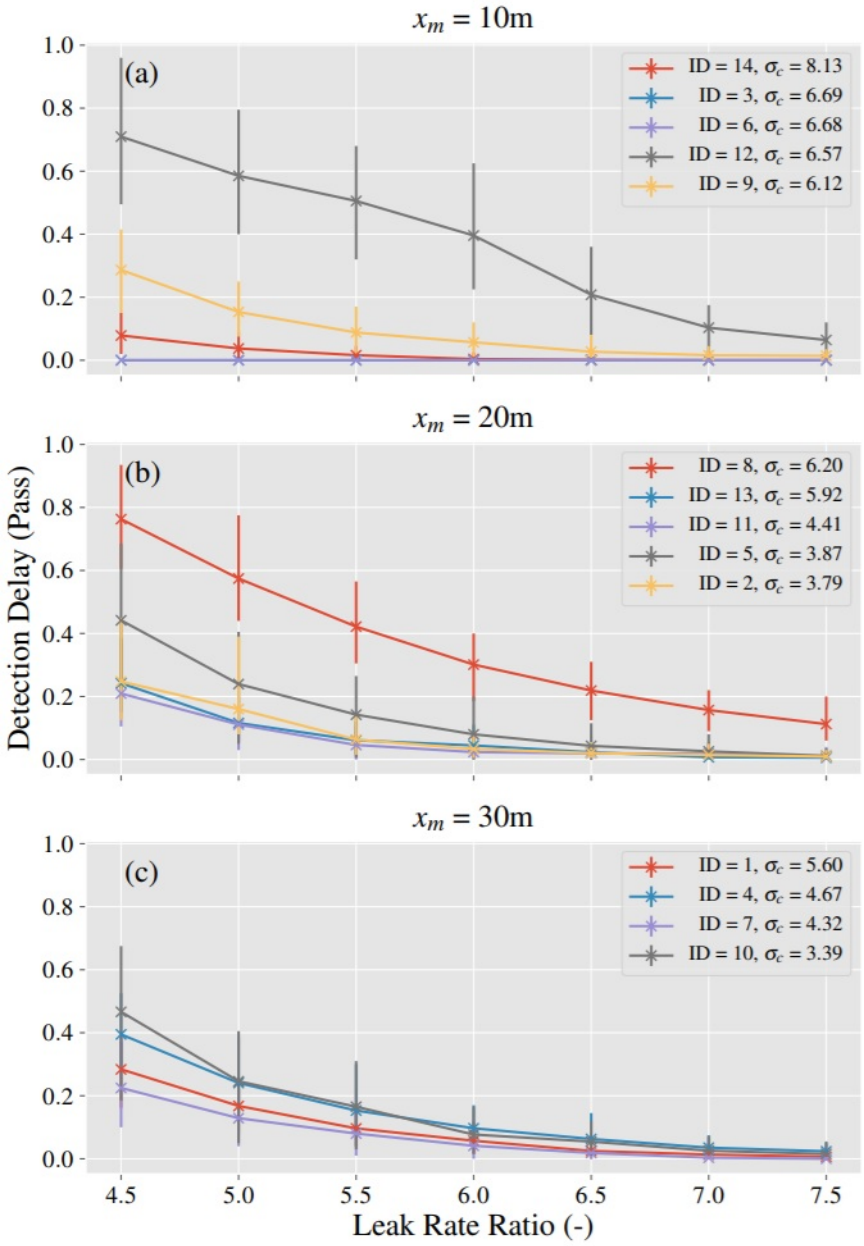

Figure 3.11: Evolution of detection delay (with units of number of passes) for a series of leak rate ratios varying between 4.5 and 7.5 for source-to-sensor distances, x_m of (a) 10m, (b) 20m and (c) 30m. ID refers to the experiment ID as seen in Table 3.1, and σ_c refers to the standard deviation of c^y measurements before the changepoint that is calculated for each experiment. Vertical bars represent the 95% confidence intervals.

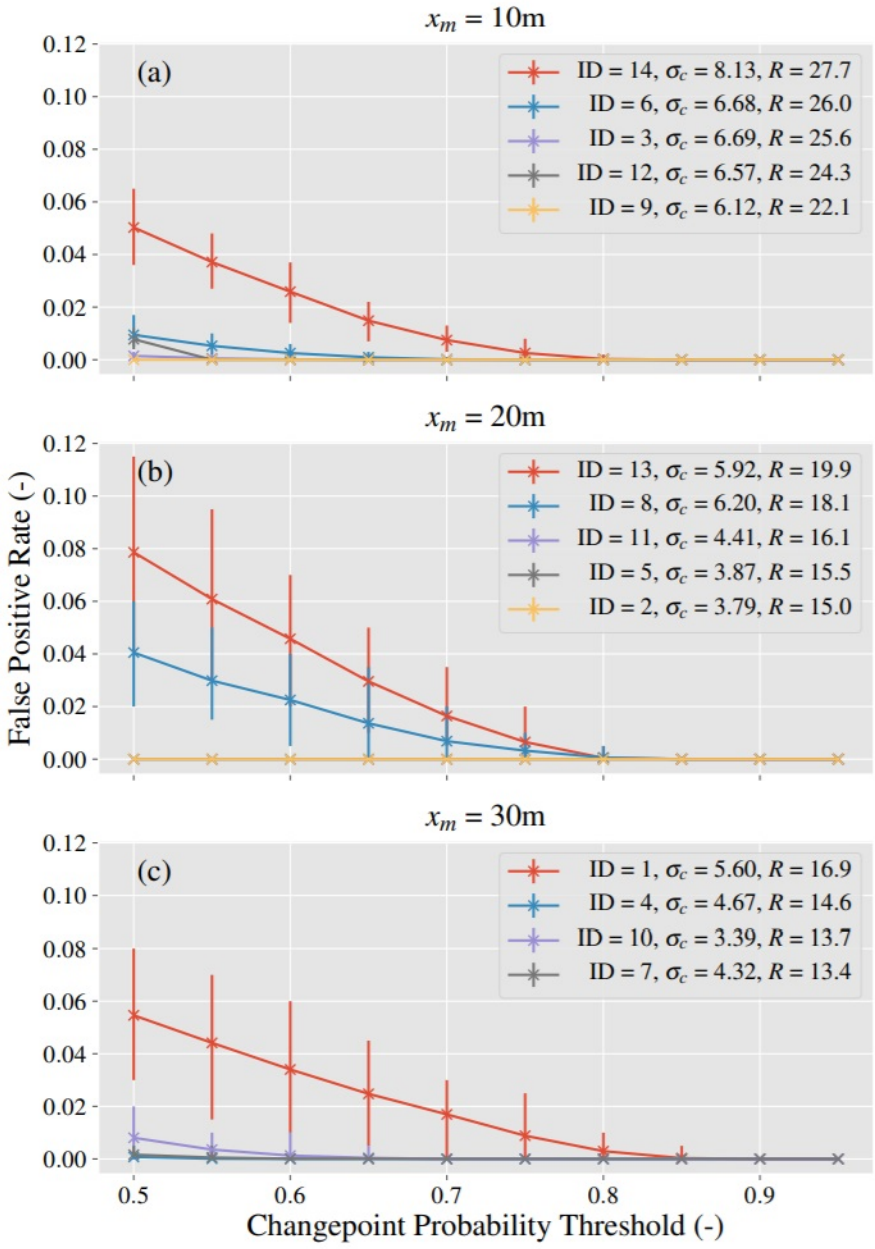

Figure 3.12: Evolution of false positive rate for changepoint probability thresholds varying between 0.5 and 0.95 for source-to-sensor distances, x_m of (a) 10m, (b) 20m and (c) 30m. ID refers to the experiment ID as seen in Table 3.1, σ_c refers to the standard deviation of c^y measurements before the changepoint, and R refers to range of c^y measurements before the changepoint for each experiment. Vertical bars represent the 95% confidence intervals.

3.5 Conclusions

In this study, we addressed the problem of detecting changes in the emission rate of a point-source by developing a recursive Bayesian scheme. This methodology directly builds on a recursive Bayesian framework that was previously used to estimate the emission rate from point sources. As a result, the introduced recursive Bayesian methodology has the ability to simultaneously detect changepoints in the emission rate and estimate the emission rate before and after changepoints. In addition, we applied our changepoint detection algorithm to a series of controlled release experiments, where a mobile sensor traversed cross-sections of the plume emitting from a point-source at different downwind distances, in the presence of an obstacle close to the source. Several measures were used to evaluate the performance of the changepoint detection methodology noting that the importance of each performance measure depends on the practical application at hand. We found that the changepoint algorithm is extremely effective (>90% success rate) in identifying changes when the emission rate is tripled. This level of success is significant given recent findings suggesting that majority of emissions from the oil and gas sector can be caused by abnormal operations that drastically and suddenly increase the emission rate (by more than an order of magnitude) [2, 3]. Further, the results showed that the statistics of the cross-plume mass concentration measurements such as the mean, standard deviation and the range can be used as predictors of the performance of the changepoint detection algorithm. Particularly, it was found that at a given leak rate ratio, lower values of coefficient of variation correspond to higher recall values which translates to higher effectiveness of the algorithm in detecting changes immediately after they occur. Moreover, it was shown that

the false positive rate of the changepoint detection algorithm was less than 2% when using a prescribed changepoint probability threshold of 0.8 for all the controlled release experiments.

Although the changepoint detection algorithm was applied to mobile sensor measurements in the near field, the methodology can be easily adapted for fenceline monitoring applications using networks of fixed sensors or far-field measurements using a single stationary sensor. In these examples, mass concentrations and meteorological conditions are often averaged over 30-minute periods. By treating each 30-minute interval similar to a single sensor pass in the experiments described in the current study, changes in the emission rate can be found using the detection algorithm.

With the changepoint detection methodology presented here applied to synthesized data from a single emission source, future work will be focused on evaluating the performance of the algorithm under more real-world scenarios such as intermittent faulty operation, and multiple emission rates caused by various operating conditions. Moreover, for practical settings, it is necessary to investigate the training time required to learn all the baseline parameters for the Bayesian inference scheme, most importantly the range of values used in the prior and the uncertainty term (σ_e) in the likelihood function of equation (3.13) before a change occurs. More studies on fault detection using advance sensing and measurement technologies will be beneficial in effective and rapid identification of large emitters which can lead to significant reductions in methane emissions from the oil and gas industry.

CHAPTER 4

A SPATIAL LAND-USE CLUSTERING FRAMEWORK FOR INVESTIGATING THE ROLE OF LAND-USE IN MEDIATING THE EFFECT OF METEOROLOGY ON URBAN AIR QUALITY

4.1 Introduction

Around the globe, exposure to air pollution causes millions of premature deaths annually [81], and is associated with chronic respiratory illnesses that increase the co-morbidity risk of many viral infections [82]. Early evidence, for example, suggests exposure to air pollution may increase mortality of COVID-19 [83]. One group of pollutants with known deleterious effects on health is Nitrogen oxides (NO_x). Nitrogen dioxide (NO_2) is commonly used as the indicator for the NO_x group and NO_2 is mainly formed by burning of fuel. Exposure to NO_2 is associated with irritation of the airways, decreased lung capacity, increased mortality from coronary heart disease, and increased incidence of diabetes, hypertension, and other cardiovascular and respiratory illnesses [84, 85]. Further, in a study of 66 administrative regions in Europe, regions with chronic exposure to NO_2 were observed to experience the highest fatality rates from COVID-19 [86]. Therefore, monitoring and mitigating exposure to NO_2 is important to public health and safety.

Traditionally, air pollution has been monitored using sparse networks of fixed stations installed in urban areas with the goal of regulatory compliance. While these fixed stations offer accurate and reliable pollutant measurements, they provide very low spatial coverage. Yet, pollutant concentrations can vary sharply over short distances due to heterogeneity in emission sources and ur-

ban form [87, 88]. In fact, it has been shown that pollutant concentrations can differ more between two neighborhoods of the same city than between two distinct cities [89]. Hence, while the networks of fixed monitoring stations remain essential for air quality regulation compliance, they fail to capture the strong spatial variability in pollutant concentrations within urban areas with strong implications for epidemiology and environmental justice [88, 90–92].

Mobile measurements show promise for overcoming the limitations of fixed-site air pollution monitoring stations [89, 93–96]. The spatial flexibility of mobile measurements has led to their application in characterizing regional pollutant concentrations and in locating pollution hotspots in select locales [93, 97–99]. While early local mobile campaigns were successful in describing spatial gradients in pollutant concentrations, many of these campaigns had limited spatial domains and were conducted for relatively short time periods. Recently, city-scale mobile monitoring campaigns have become more common [88,94,100,101], with vehicles outfitted with state-of-the-art sensors and deployed to cover extensive parts of urban areas over several months and years, allowing for repeated sampling of visited locations. Repeated sampling coupled with data analytics algorithms grants statistical power to construct stable, long-term spatial maps of pollutant concentrations at high resolutions over large areas [88,94,100]. These spatial maps are useful in depicting persistent patterns in pollutant concentrations, measuring average pollution (averaged over a year) in a region, and locating air pollution hotspots. However, temporal variability in air pollution is typically not reported, despite its vital importance for identifying the time of exposure above key concentration thresholds of human health significance [82].

Temporal dynamics of pollutant concentrations within an urban area are de-

pendent on both the regional (city-wide) meteorology for overall atmospheric boundary layer mixing and the local meteorology, as modulated by local urban form, for its control on ground level concentrations. In other words, local land use affects the temporal dynamics of air quality by mediating the relationship between regional and local meteorology (i.e. some areas more or less venitlated than others). Meanwhile, the effects of regional meteorology on air quality are known to vary between seasons [102, 103]. Therefore, the study of the temporal variability of pollutant concentrations requires local pollutant measurements over different seasons as done in large scale air quality measurement campaigns. One such campaign was the mobile measurement effort by two Google Street View Cars in Oakland, CA, sampling ambient NO_2 concentrations with a frequency of 1-Hz over a two-year period. This novel dataset provides information on pollutant concentrations of all city streets within the study domain of West Oakland, downtown Oakland, and East Oakland across different seasons and under varying meteorological conditions.

In this chapter, we investigate the role of urban land form in mediating the effect of regional meteorology on intra-urban air quality in Oakland, CA using the Google Street View air quality dataset. To the best of our knowledge, this is the first study focusing on using city-wide mobile measurements to examine spatially varying temporal patterns in air quality due to interaction between meteorology and urban form. To this end, we developed a data-driven spatio-temporal framework as follows. First, inspired by findings of Messier et al. (2018), we clustered the spatial locations in Oakland, CA based on land-use covariates (as surrogates for emission sources and urban form) using the k-means clustering algorithm [104, 105]. This clustering effectively reduces the spatial fidelity of the data, but increases its statistical power by producing clusters with

large sample sizes. The increase in statistical power is required for successful data stratification based on wind speed and season. Subsequently, we used conditional averaging to characterize the effect of wind speed on NO_2 concentrations in each cluster. We note that the focus on wind speed as an effective temporal variable in modulating NO_2 concentrations and the need for clusters with large sample sizes are discussed in detail in our exploratory analysis described in section 4.3. The analysis is concluded with the study of exceedance probabilities under varying seasons and wind speed conditions. Exceedance probabilities are an important measure of exposure to extreme pollutant concentrations, with clear ties to acute effects of air pollution on human health. The main contribution of this chapter is providing a framework that exploits land-use variables to learn about the relationship between meteorology and intra-urban air quality using limited air pollution data from mobile sensors. The second contribution is the development of a land-use clustering technique consisting of the k-means algorithm and a comprehensive procedure for selecting the number of clusters. The third contribution is the application of the framework to pre-existing data from Oakland, CA and the insightful results related to how urban form modulates the effect of wind speed on intra-urban air quality.

4.2 Data

Multiple datasets including meteorological data, land-use data and mobile NO_2 measurements, were analyzed in this study to investigate the effect of meteorology and land use on air pollution levels in distinct regions of Oakland, CA, with use cases of each dataset presented in Figure 4.1. In this figure, dark blue rectangles correspond to data and orange rectangles correspond to data processing

steps. Further, variables in bold refer to data matrices and non-bold variables indicate vectors, and light blue rectangles refer to the methodologies used in this study.

4.2.1 Data sources

Mobile measurements of 1-Hz NO_2 concentrations were collected in Oakland, CA in a joint effort between University of Texas at Austin, Aclima Inc., Google and the Environmental Defense Fund, details of which are available in [88]. In brief, Aclima environmental intelligence fast-response pollution measurement and data integration platforms were installed on two Google Street View mapping vehicles. These vehicles measured weekday daytime NO_2 concentrations on city streets. Measurements were collected on every road in a 30 km^2 domain, incorporating residential, commercial and industrial areas [106]. The data includes more than 2.7 million samples from two datasets with measurements from a total of 305 days from July 13, 2015 to August 31, 2017. Our data reduction "snapping" scheme follows that of Messier et al. [104]. First, we divided a street centerline file (obtained from OpenStreetMaps.com) into more than 19,000 30-meter road segments. Next, we employed a nearest-neighbor algorithm (Python SciPy "ckdnearest" algorithm) to "snap" each 1-Hz measurement to its nearest road segment resulting in consistently defined locations [107]. The snapping procedure is described in more detail in Appendix B.1.

We also gathered land-use data for each 30 meter road segment in the form of 26 binary and continuous geographic covariates following the methods of Messier et al. [104]. The procedures used for calculating the geographic covari-

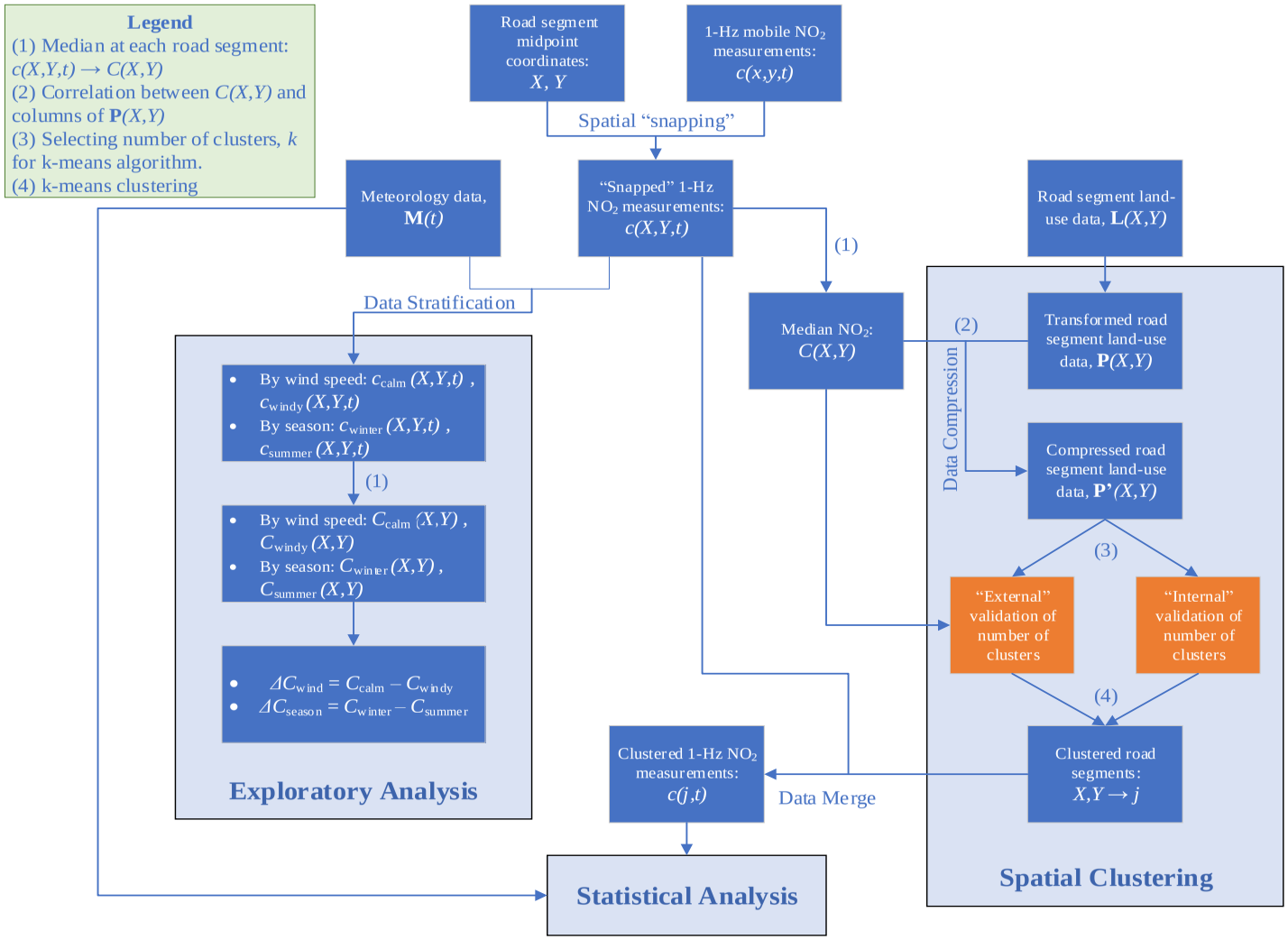

Figure 4.1: Flow diagram depicting the evolution of the data.

ates are described in detail in Appendix B.2 with the full list of the covariates presented in Table B.1.

Surface meteorological observations, including hourly temperature, wind speed and direction, and precipitation, for Oakland International Airport were acquired from the National Oceanic and Atmospheric Administration (NOAA) Automated Surface Observing Systems (ASOS) through Iowa Environment Mesonet (IEM) portal maintained by Iowa State University (`https://mesonet.agron.iastate.edu/`; accessed November 1, 2020). A major strength of the ASOS is the consistency of measurements in reporting wind data which is a crucial variable in this study. Hourly solar radiation data in the form of Global Horizontal Irradiance (GHI) was obtained from Solcast (`https://solcast.com/`; accessed November 5, 2020) using the Solcast API, a source for satellite-derived solar irradiance data. The data was obtained for a location of 37°48′54″N 122°16′57″W and is within the core of the West Oakland/Downtown domain of NO_2 mobile measurements and coincides with the fixed-site regulatory monitor located at the Oakland West site managed by the Bay Area Air Quality Management District (BAAQMD). We then used linear interpolation to convert the observations to match the 1-Hz measurements of the mobile campaign, therefore augmenting the NO_2 observations with surface meteorological measurements and satellite-driven radiation measurements.

4.2.2 Selection of temporal variables

NO_2 concentrations in urban areas are affected by regional meteorological variables. Strong inter-dependencies between different meteorological variables,

complicate the relationship between these variables and pollutant concentrations. Establishing links between regional meteorology and pollutant concentrations is further complicated by the role of local urban land form in mediating the effect of regional meteorology on the local mixing within the urban area. Therefore, prior to our statistical analysis, we apply a variable selection procedure driven by the regional meteorological conditions during the measurement period and unique to the study area of Oakland, CA.

The climate in Oakland is characterized by dry, warm summers and mild, wet winters. However, during the measurement campaign precipitation data was reflective of prevailing drought conditions (zero precipitation for more than 99% of all study hours). In addition, the prevailing wind direction was found to be from the West for approximately 85% of all study hours. Due to the low variability observed in wind direction and precipitation during the study period, the effects of these parameters on intra-urban NO_2 pollution are not pursued here.

While high daily temperatures have been previously linked to higher concentrations of NO_2, increases in global radiation have been shown to correlate with reduced NO_2 concentrations [102]. The lack of nighttime measurements coupled with a moderate positive correlation (Spearman's correlation coefficient = 0.57) observed between temperature and radiation during the study period, leads to the conclusion that isolating the effect of each of these variables is not viable in our analysis. On the other hand, pollutant concentrations, including NO_2, are known to be seasonal [102, 108]. Henceforth, we assume that investigating the seasonality in the data indirectly accounts for the effects of emission seasonality, temperature and radiation. Therefore, temperature and

radiation are excluded from the analysis and instead a seasonal stratification of concentration data as described in section 4.3 is adopted.

A secondary variable with known effects on atmospheric dispersion that can be calculated from the available data (radiation and wind speed) is atmospheric stability. In urban areas however, the increased drag force caused by roughness obstacles (e.g. buildings and other structures) leads to larger friction velocities than in open areas. Therefore, stability over urban areas is biased towards neutral (adiabatic) conditions [109]. As a result, the effects of atmospheric stability on intra-urban air pollution are not pursued in this study, due to low variability in stability conditions.

In this study, we primarily investigate the effects of wind speed on intra-urban NO_2 concentrations, as it has been established as an important meteorological parameter in affecting NO_2 pollution by previous studies [102, 110, 111]. In addition, seasonality of NO_2 concentrations in Oakland are studied. The exploratory analysis in section 4.3 further validates the choice of wind speed as an important meteorological parameter controlling NO_2 concentrations across the city of Oakland.

4.3 Exploratory data analysis

Prior to clustering, we conducted a preliminary analysis to examine the relationship between the selected variables in section 2.2 and NO_2 observations on 30-m road segments. The analysis relies on data stratification which refers to partitioning the concentration data into distinct and non overlapping groups of independent variable states. Two distinct stratifications are applied to the

data separately to identify effects of wind speed and seasonal changes on NO_2 concentrations, respectively. Wind speed stratification is carried out by dividing all 1-Hz NO_2 measurements into two groups: wind speeds below 3.5 m/s (calm) and above 5.5 m/s (windy). The threshold values of 3.5 and 5.5 m/s are chosen for the following reasons: 1) similar sample sizes between the two groups, and 2) a wind speed buffer of 2 m/s prevents misclassification as the accuracy of the ASOS monitoring system is 1 m/s. After stratification, each group is analyzed separately to calculate the median of 1-Hz NO_2 measurements (C_{calm}, C_{windy}) for those 30-m road segments that have been visited on at least 10 distinct days, noting that 10 distinct measurement days ensure stable estimations of median concentrations [88]. Lastly, the local differences in median NO_2 concentrations (ΔC_{wind}) between calm and windy measurements are computed as $\Delta C_{wind} = C_{calm} - C_{windy}$ for each 30-m road segment (Figure 4.2a). The spatial distribution shows the contrast between the median concentrations, with the mean (median) ± standard deviation of ΔC_{wind} = 8.0 (7.6) ± 5.8 ppb. It is worth noting that an increase of 5.3 ppb in long-term NO_2 concentrations (averaged over one year or more) has been associated with all-cause mortality with hazard ratios of 1.01 – 1.03 (95% CI), highlighting the significance of the computed ΔC_{wind} [112].

Seasonal stratification is carried out by dividing the 1-Hz NO_2 measurements into two groups: November 1st until February 28th are labeled winter measurements and May 1st until August 31st are labeled summer. Following similar steps as the wind speed analysis, the local differences in median NO_2 concentrations (ΔC_{season}) between winter and summer are computed as $\Delta C_{season} = C_{winter} - C_{summer}$ (Figure 4.2b) for each 30-m road segment. The spatial distribution of ΔC_{season} indicates higher median concentrations during winter

which is in agreement with our analysis of hourly NO_2 observations from the fixed site monitoring site in West Oakland (Figure B.7 in Appendix B.3). The mean (median) ± standard deviation of ΔC_{season} is 8.0 (7.1) ± 5.1 ppb.

Our exploratory analysis reveals the effect of wind speed on NO_2 concentrations through a two-group stratification (windy and calm), because a multi-group stratification would not be appropriate as very few road segments would pass the 10 distinct day selection criterion. Furthermore, a mixed stratification based on wind speeds and seasons leading to 4 groups (e.g. winter and windy, summer and calm, etc.) would not be viable for the same reason. Therefore, we propose an approach that uses cluster analysis to group together road segments that are similar in terms of land use to investigate the effect of each temporal control separately and with finer granularity (i.e. more wind speed intervals). This clustering approach increases the statistical power of our temporal analysis, because of significantly larger sample sizes of each cluster compared to individual road segments.

4.4 Methodology

4.4.1 Spatial clustering

A popular approach for quantifying intra-urban variation in air pollution is land-use regression (LUR) [113–115]. LUR models are mainly used to depict spatial variation of air pollution and do not give any information on temporal variations of air quality. Furthermore, time series analysis of the mobile measurements is not feasible as the data are collected along spatio-temporal paths

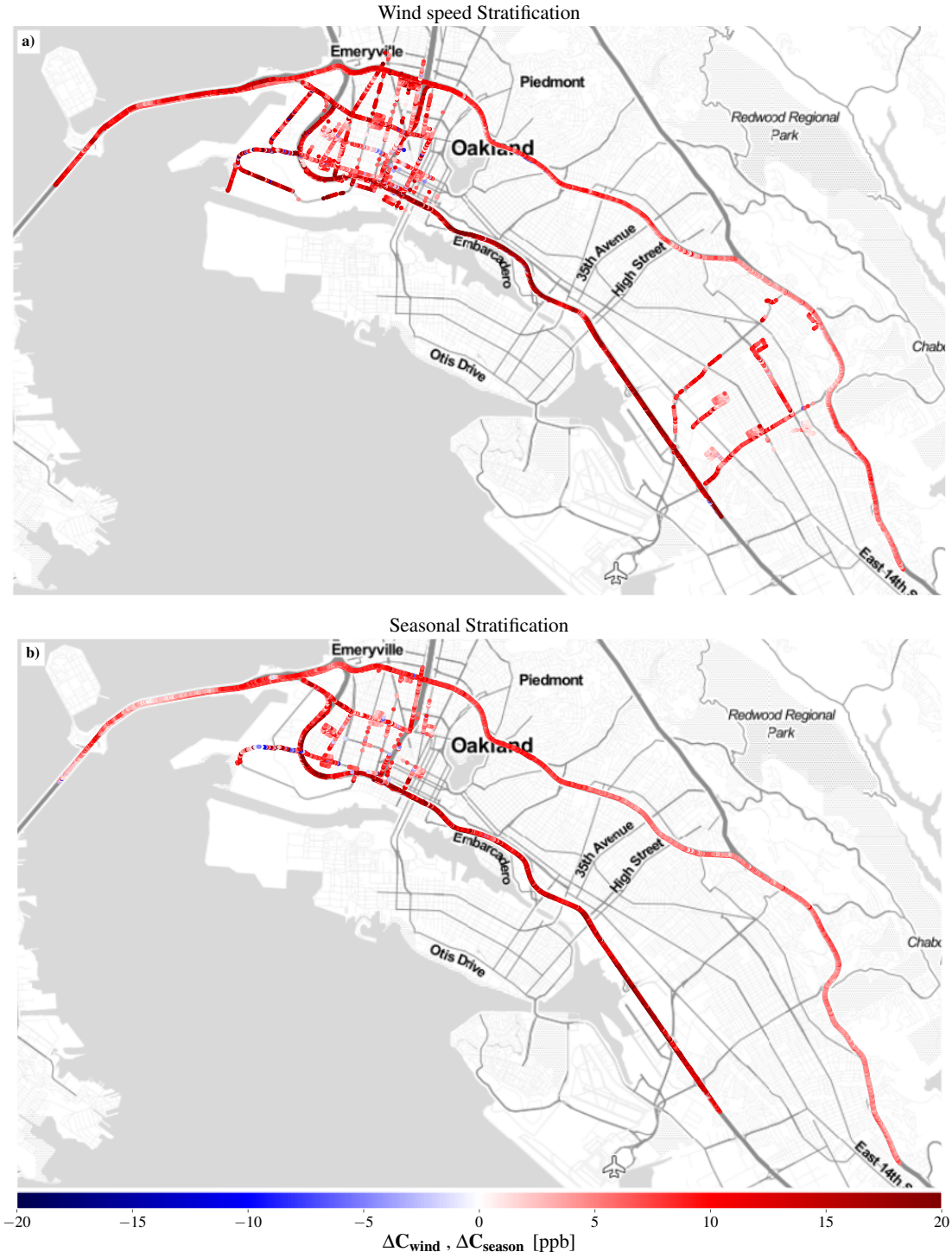

Figure 4.2: Difference in median NO_2 concentrations between (a) calm and windy and (b) winter and summer observations. Map tiles by Stamen Design. Map data by OpenStreetMap.

(cars traversing the city). In addition, the size of the dataset is inadequate to resolve the effects of all the factors influencing pollutant concentrations at every 30-m road segment.

Inspired by LUR models which suggest that locations with similar land use characteristics have similar pollutant concentrations, we aim to overcome the sample size issue by clustering the 30-m road segments based on their land-use covariates, and then study the temporal evolution of NO_2 concentrations within each cluster. This allows us to examine how land use modulates the effect of regional meteorology on local air quality dynamics.

Clustering is an unsupervised learning method for grouping a set of objects in a way that objects in the same group are more similar to each other than to those in other groups. The similarity of objects is assigned by the features that clustering is based on. In this study, we cluster 30-m road segments in the city of Oakland, CA, by using land-use covariates of these road segments as features. As discussed in section 4.2.1 a total of 26 land-use covariates are considered. Furthermore, it is desirable that road segments that are geographically close to each other fall in the same cluster, as we expect the effects of emission sources and local meteorology to be similar for adjacent road segments. Therefore, the latitude and longitude of the center point of individual road segments are also included as features in the clustering algorithm bringing the total feature count to 28.

Data pre-processing

Performance of clustering algorithms are generally improved when the number of features are lowered [116]. First, we lower the number of features using a principal component analysis (PCA). Feature reduction using PCA is appropriate in the land use context, because the land-use variables considered are highly correlated with each other, containing redundant information that is detrimental to the performance of clustering algorithms. Prior to PCA, the features are standardized by subtracting the feature mean and rescaling the feature variance to unity. The standardized features are then stored in an $n \times 28$ matrix, with n being the number of unique road segments. Performing PCA on this preliminary matrix leads to a new $n \times 28$ matrix that we label matrix $\mathbf{P}$:

$$\mathbf{P} = (\mathbf{p_1}, \mathbf{p_2}, \ldots, \mathbf{p_{28}}) = \begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,28} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,28} \\ \vdots & \vdots & \vdots & \vdots \\ p_{n,1} & p_{n,2} & \cdots & p_{n,28} \end{bmatrix}, \tag{4.1}$$

where each column vector $\mathbf{p_j}$ corresponds to the newly formed principal components (PCs) that are linearly uncorrelated with each other. The PCs are ordered based on amount of variance in the original variables accounted for by each component, with PC_1 accounting for the most variance and PC_{28} accounting for the least. The first 12 PCs account for approximately 85% of the variance in land-use variables. A more detailed description of the PCA is presented in Appendix B.4.

To further reduce the number of features, out of the first 12 PCs, we retain those PCs that are correlated with median NO_2 concentrations computed for each road segment. Therefore, we calculate the Pearson correlation coefficients

of the columns of **P** and the column vector **C**:

$$\mathbf{C} = (C_1, C_2, \ldots, C_n)^T , \tag{4.2}$$

with C_i computed as median of NO_2 concentration at road segment i. Labeling the Pearson correlation coefficient between $\mathbf{p_i}$ and **C** as ρ_i, we only retain those PCs that satisfy $|\rho_i| > 0.1$. This analysis results in the retainment of 4 PCs that account for approximately 60% of the variance, therefore, greatly reducing the number of features prior to clustering.

Clustering method

We apply the k-means algorithm developed by Hartigan and Wong (1979) to cluster the 30-m road segments [105]. This algorithm seeks to partition n points (30-m road segments) in D dimensions (4 PCs in this case) into k clusters. It iteratively searches for a local solution that minimizes Euclidean distance between the points and cluster centers. The initial cluster centers in the k-means algorithm can be chosen randomly, by the user or by randomized techniques. Here, we utilize the popular "k-means++" initializing algorithm as it seeks to spread out the cluster centers, a desirable property in this study [117]. The main advantages of k-means are its ease of implementation, computational efficiency, and reduced sensitivity to outliers compared to hierarchical clustering methods.

Selecting the number of clusters

In k-means clustering the main required hyper parameter is the number of clusters (k) which is often not known a priori. The number of clusters can be assigned by either pre-existing knowledge of the data that is not available from

the dataset itself, or by providing a descriptive statistic for ascertaining the extent to which the observations comprising the dataset fall into natural distinct groupings [118]. In short, the number of clusters can either be assigned solely through the dataset (Data-based or internal methods) or by additional knowledge obtained externally (External methods). In this study, we apply both internal and external methods to select the optimal value of k and validate the clustering analysis. To select the number of clusters, clustering solutions are first found for a sequence of consecutive k values between 5 and 15. These solutions are then compared to each other using internal and external methods to find the optimal number of clusters.

Internal method The gap statistic approach originally introduced by Tibshirani et al. is among the standard data-based methods for choosing the number of clusters in a dataset [119]. This method utilizes the total "within-cluster dispersion", which is defined as the sum of the distance between each data point (road segment features) in the cluster and the cluster center. For each value of k, the k-means algorithm is applied to the observed data and a randomly generated data set that uniformly spans the feature space and has the same size as the observed data. The gap function, Gap(k), is then computed as the difference between the sum of the total within-cluster dispersion for the observed and random data (generated 100 times in this analysis). The optimal number of clusters for the given data set is the smallest k such that

$$\text{Gap}(k) \geq \text{Gap}(k+1) - s_{k+1}, \tag{4.3}$$

where s_{k+1} is the standard deviation of the total within-cluster sum of squares of the randomly generated data.

External method In regards to applying additional knowledge to assign the number of clusters, we consider the cluster average of variability of median NO_2 concentration for all 30-m road segments within each cluster. Variability labeled V, is calculated as the standard deviation from the mean of median daytime concentrations for 30-m road segments within each cluster:

$$V(j) = \left(\frac{1}{n_j} \sum_{i=1}^{n_j} \left(C_i^{(j)} - \bar{C}^{(j)} \right)^2 . \right)^{1/2} \tag{4.4}$$

where n_j is the number of road segments in cluster j, $C_i^{(j)}$ is the median NO_2 concentration observed at the i'th road segment belonging to cluster j and $\bar{C}^{(j)}$ is the mean of median NO_2 concentrations observed at all road segments belonging to cluster j. Average cluster variability, labeled S, is then calculated as follows:

$$S(k) = \frac{1}{k} \sum_{j=1}^{k} V(j) \tag{4.5}$$

At first glance, solutions with lower average variability may be judged to be superior to those with higher average variability. However, average variability within clusters generally tends to decrease with increasing number of clusters. Therefore, we create a "benchmark" for every value of k, and judge the superiority of solutions based on their distance from this benchmark. For each k, the benchmark is created by first sorting 30-m road segments by their corresponding value of median NO_2 concentrations and then grouping the road segments into k equally-sized clusters. We then find the number of clusters that minimizes the difference between average variability of the median NO_2 concentrations of the original clustering using k-means algorithm, $S(k)$ from Eq. 4.5, and the average variability of median concentrations of the benchmark, $S^*(k)$, for k values between 5 and 15:

$$\underset{k \in [5,15]}{\arg\min} \left[S(k) - S^*(k) \right] . \tag{4.6}$$

4.4.2 Statistical analysis

Once the road segments are clustered, the effects of wind speed and seasonality on 1-Hz NO_2 concentrations corresponding to road segments in each cluster are investigated. Similar to section 4.3, NO_2 concentrations in each cluster are stratified into two groups based on the measurement season. Following this division, conditional averaging based on wind speed is employed to quantify the effect of wind speed on NO_2 concentrations for each cluster/season combination. Further, probabilities of NO_2 exceeding pre-determined thresholds are calculated through a two-step sampling process for every cluster, season and wind speed condition.

Conditionally averaged concentration

Every NO_2 concentration measurement coincides with a wind speed measurement as described in section 4.2.1. The concentration values are organized based on the wind speed such that multiple concentration values are grouped together within a given wind speed interval, U. The conditionally averaged NO_2 concentration value, denoted $\langle c|u \rangle$, is calculated within designated wind speed intervals as shown:

$$\langle c|u \rangle = \frac{1}{N_U} \sum_{u_i \in U(u)} c\left(u_i\right), \tag{4.7}$$

where c represents 1-Hz NO_2 measurements, $U(u) = \{u_i : -\Delta u/2 \leq u - u_i < \Delta u/2, \forall i = 1, 2, \ldots, N_U\}$ and N_U is the total number of data points within the given wind speed interval U. In this analysis, Δu is set to 1 m/s. This choice of the wind speed intervals is driven by the accuracy of 1 m/s of the ASOS monitoring system and the available sample size of NO_2 measurements coinciding

with each given interval. In addition, conditional probability distribution functions (PDFs) of concentration are also constructed to calculate the conditional interquartile range in a similar manner to the conditional averages.

Exceedance probabilities

Exceedance probabilities are calculated by computing empirical cumulative distribution functions (ECDFs) of NO_2 concentrations for every cluster, season and wind speed condition. Due to the streaming nature of mobile measurements, observations recorded on any given day are correlated, particularly if the observations were recorded over a short period of time (e.g. one hour). Furthermore, the number of measurements on each day varies widely across different days, especially after cluster, season and wind speed stratifications. Therefore, direct calculation of the ECDFs using raw 1-Hz measurements gives extra weight to days with high number of measurements and biases calculated exceedance probabilities. To overcome this issue, we utilize the following two-step sampling strategy to compute ECDFs and exceedance probabilities. For each cluster, season and wind speed condition, the steps are as follows:

1. Randomly select a day with replacement from the days with at least 100 mobile measurements for the given cluster, season and wind condition.
2. Randomly sample $N = 100$ NO_2 measurements with replacement from the selected day.
3. Repeat the first two steps $N_D = 10$ times to create an ECDF with $N_D \times N = 1000$ samples.
4. Calculate exceedance probability as: $\mathbb{P}_E(T)$ = (**Number of samples with**

concentrations $> T)/(N_D \times N)$.

with T corresponding to the concentrations threshold chosen for NO_2. A robust estimate of the exceedance probability is then computed by repeating the steps above 1000 times to account for variability introduced through the random selection. We note that the data corresponding to days with less than 100 measurements account for less than 5% of all the data for a given cluster, season and wind condition, and therefore unlikely to have a significant effect on the calculated probabilities. In addition, $N_D = 10$ is chosen since there are at least 10 unique measurement days with at least 100 measurements for each cluster, wind and season condition.

4.5 Results and Discussion

4.5.1 Spatial clustering

After pre-processing the land-use data corresponding to individual 30-m road segments described in section 4.4.1, we select the number of clusters k, using both data-based and external methods.

Internal method We computed the gap statistic for clustering solutions between 5 and 15 clusters to find the optimal number of clusters suggested by this method. The gap statistic for these solutions are shown in Figure 4.3a with the vertical error bars corresponding to the standard error, s_k. Based on equation 4.3, this method assigns 7 clusters as the optimal value for k.

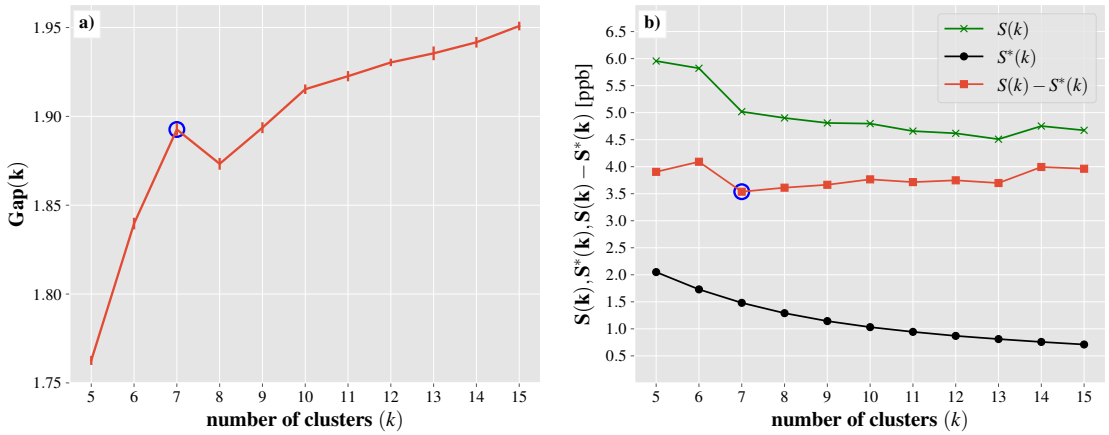

Figure 4.3: Selecting optimal number of clusters through (a) gap-statistic as an internal method suggesting 7 clusters as the optimal choice for k, with the vertical lines corresponding to s_k and (b) comparison of average within-cluster variability of daytime median NO_2 concentrations between the clustering solution and clustering benchmark as an external method, suggesting 7 clusters.

External method As discussed in section 4.4.1, we computed the statistics required to select the optimal number of clusters k using information external to land-use and location data. The results are shown in figure 4.3b where $S(k)$, $S^*(k)$ and their differences are plotted for clustering solutions between 5 and 15 clusters. Since the goal is to minimize $S(k) - S^*(k)$, this methodology indicates that the optimal choice for k is 7 clusters.

Since both validation methods yield the same result regarding the optimal number of clusters, 7 was chosen as the number of clusters. Figure 4.4a below shows the clustering solution utilizing the k-means algorithm with $k = 7$ as a spatial map of Oakland, CA. Meanwhile, Figure 4.4b presents the histograms of median NO_2 concentrations at each road segment belonging to each of the 7 clusters. This clustering solution shows that cluster 1 is a mixture of highways and major roads in industrial areas closer to East Oakland, cluster 2 covers resi-

dential areas in East Oakland that are located at higher elevations (>100m higher than sea level), cluster 3 mostly includes both major and narrow roads in industrial zones of West Oakland and Downtown, cluster 4 covers highways that are truck prohibited, cluster 5 mostly covers residential zones and roads located in East Oakland, cluster 6 mostly consists of interstate highways that allow truck passage and cluster 7 covers residential areas in West Oakland and Downtown. Based on these findings, the clusters will be referred to using the following labels:

- Cluster 1 - Industrial East Oakland
- Cluster 2 - Elevated residential East Oakland
- Cluster 3 - Industrial West Oakland
- Cluster 4 - Truck prohibited highways
- Cluster 5 - Residential East Oakland
- Cluster 6 - Truck-route highways
- Cluster 7 - Residential West Oakland

With geographically similar road segments grouped together in clusters with a significant number of mobile NO_2 measurements available within each cluster, mobile measurements within each cluster can be investigated with regards to wind speed and seasonal changes.

4.5.2 Effects of wind speed on concentrations

For each cluster, effects of wind speed on NO_2 concentrations during each season are examined through conditionally averaged concentrations and are

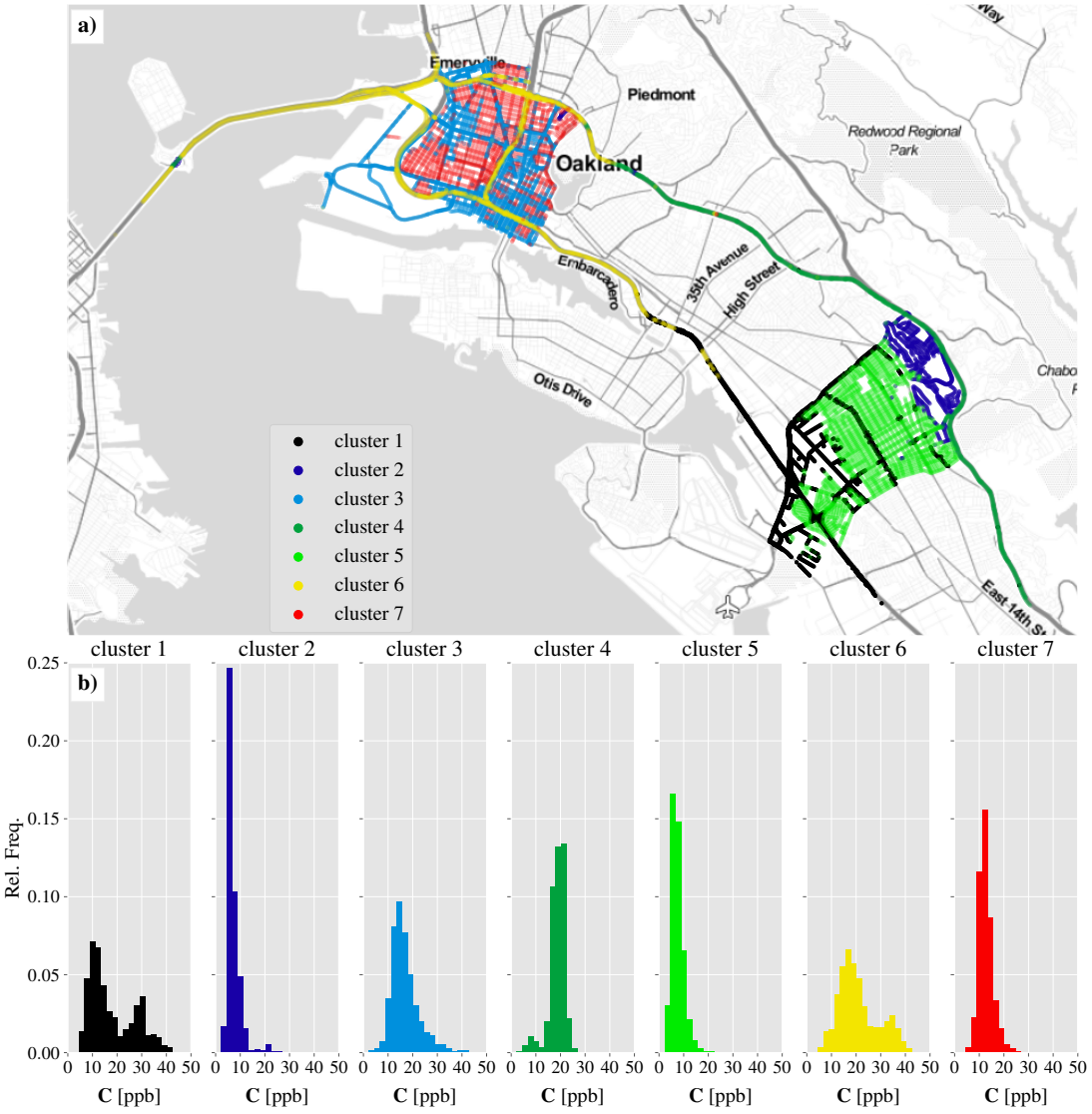

Figure 4.4: Clustering 30-m road segments into $k = 7$ clusters. (a) Spatial map of 30-m road segments, color coded based on cluster numbers, and (b) histograms of daytime median NO_2 concentrations for each cluster. Map tiles by Stamen Design. Map data by OpenStreetMap.

shown in Figures 4.5 and 4.6 for winter and summer, respectively. The results are shown for 4 of the 7 clusters including Industrial and residential West Oakland and inter-state highways (i.e. clusters 3, 4, 6 and 7) for the following reasons: 1) These regions cover highways, industrial and residential zones where the population lives, works and commutes, 2) the results allow for comparisons between residential/industrial zones, truck-route/truck-prohibited highways, and highway/non-highway roads, and 3) the majority of mobile measurements were made in these regions and therefore sample sizes are large enough for statistically significant analyses.

During winter, the West Oakland clusters follow a similar downward trend as measured by a linear fit to the conditionally averaged concentrations, even though concentrations are generally higher in the industrial cluster. While the concentrations on truck-route highways also drop with increasing wind speed, the drop is smaller than West Oakland. A plausible explanation for this behaviour is the additional turbulence on the highways caused by moving traffic which increases vertical mixing of the pollutants with the clean air above even in the absence of wind. This additional turbulence in turn leads to a smaller marginal effect of wind speed on NO_2 concentrations. This hypothesis can be validated by comparing the average velocity of the sensing vehicle traveling in each of the clusters (Table 4.1) which is used as a proxy for traffic density. Concentrations on truck prohibited highways do not follow a significant downward trend which is likely due to traffic turbulence (highest average car speed among clusters) and the topography of this cluster, located at higher elevations compared to other investigated clusters.

In the summer, the conditionally averaged concentrations do not follow a

Table 4.1: The average speed of the Google Street View Car (sensing vehicle) in units of [m/s] within each cluster during Winter and Summer.

Season	Industrial West Oakland	Residential West Oakland	Truck-Prohibited Highways	Truck Route Highways
Winter	11.3	8.9	30.9	24.7
Summer	8.7	7.1	28.4	22.5

significant trend in any of the clusters, suggesting that wind speed is a less important predictor of NO_2 concentrations compared to winter. One possible explanation for this behaviour is increased vertical mixing in the summer caused by increased radiation and surface heat fluxes that leads to overall lower concentrations in the summer. It is worth noting that the concentrations observed for each cluster during summer is consistently lower than those observed in the winter, as evident through a comparison between figures 4.5 and 4.6 which is in agreement with the exploratory analysis of section 4.3. Moreover, the slightly positive slope observed for the truck-prohibited highways cluster can be attributed to the travel of pollutants from West Oakland due to Westerly winds.

4.5.3 Exceedance probabilities

For each cluster, the probability of observing NO_2 concentrations above the threshold of 40 ppb (95th percentile of concentrations observed for the investigated clusters) are calculated under four conditions based on wind speed and seasonality as depicted in Figure 4.7. The four conditions are obtained through a mixed data stratification process following the steps described in section 4.3. The truck-route highways cluster shows a sharp drop in exceedance probabilities during windy conditions compared to calm conditions with a 53% drop

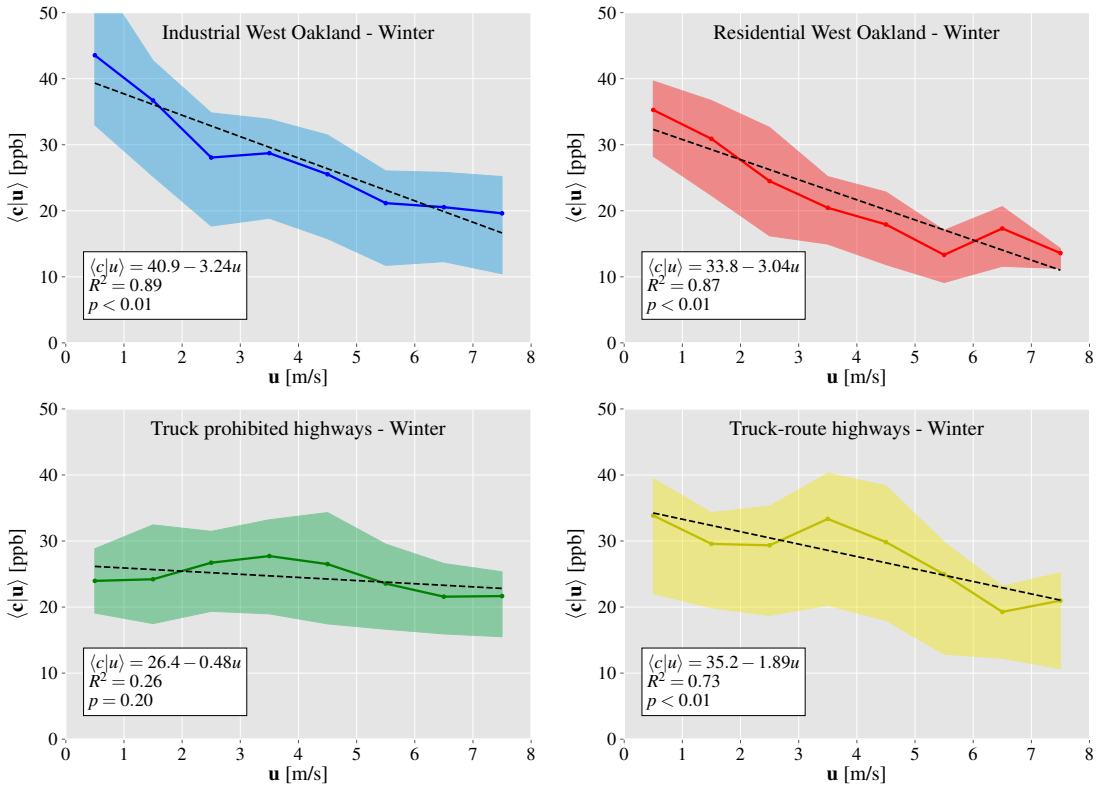

Figure 4.5: Effect of wind speed on NO_2 concentrations for each cluster during Winter. The colored solid lines correspond to conditionally averaged concentrations found through Eq. 4.7. Shaded regions correspond to the interquartile range of conditional concentration distributions. The black dashed lines correspond to a linear fit to the curve with details of the fit described in the text boxes, where coefficient of determination is represented by R^2 and the significance of the slope of the linear fit is quantified through t-tests with the p-values shown.

during winter and a 84% drop in the summer. One possible explanation for this sharp drop is tied to traffic density and speed of cars on the highway. In particular, for the truck-route highway cluster the average velocity of the sensing vehicle when NO_2 concentrations were above 40 ppb was found to be 16.7 and 15.3 m/s during Winter and Summer, respectively. Note that these values are well below the velocities shown in Table 4.1. Considering that high NO_2 are often due to high traffic during which cars are moving slowly, therefore not

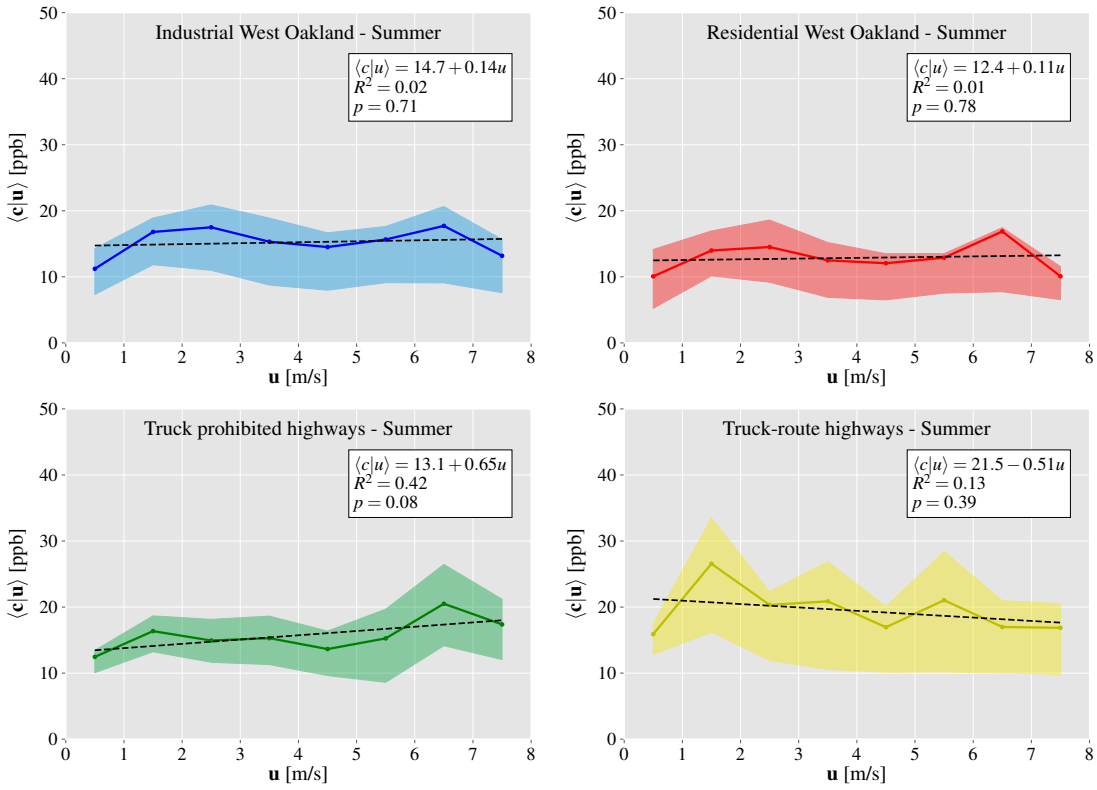

Figure 4.6: Effect of wind speed on NO_2 concentrations for each cluster during Summer. As in Figure 4.5, the colored solid lines correspond to conditionally averaged concentrations found through Eq. 4.7. Shaded regions correspond to the interquartile range of conditional concentration distributions. The black dashed lines correspond to a linear fit to the curve with details of the fit described in the text boxes, where coefficient of determination is represented by R^2 and the significance of the slope of the linear fit is quantified through t-tests with the p-values shown.

contributing to turbulence and mixing of the pollutants. In these conditions wind can be effective in creating additional turbulence that leads to the mixing of the pollutants and lowers pollutant concentrations. The significant difference between the probabilities of the two highway clusters highlights the effect of trucks and high emitting vehicles on high NO_2 concentrations. In addition, almost all of the measurements on truck prohibited highways during summer fall below the 40 ppb threshold, leading to very small exceedance probabilities. The

trend observed for the industrial West Oakland cluster is similar to that found in section 4.5.2, with exceedance probability dropping under windy conditions and lower values observed during summer. Moreover, there is a perceptible difference between the two West Oakland clusters, highlighting the correlation between land use and pollutant concentrations.

It is worth noting that the 40 ppb threshold is smaller than regulatory limits for short term exposure. Nevertheless, the exceedance probability analysis was worthwhile as it showed that the response of the tails of the concentration distribution to wind speed differed from the response of the mean concentrations. Furthermore, NO_2 levels are correlated with other pollutant concentrations highlighting the importance of an exceedance probability analysis in the context of exposure to other air pollutants in addition to NO_2 [120].

4.6 Sensitivity Analysis

4.6.1 Sensitivity of wind effects to wind speed intervals

The linear fits to the conditionally averaged concentrations found in Section 4.5.2 are subject to the chosen wind speed intervals. As such we repeated the analysis to compute the slope of the linear fit to the conditionally averaged concentrations for different lengths of the wind speed intervals, Δu, varying between $0.5 m/s$ and $1.5 m/s$. The calculated slopes for different wind speed intervals for each cluster during winter are provided in Table 4.2, indicating that the magnitude of the calculated slopes depend on the wind speed intervals. Nevertheless, these results confirm that the effects of wind speed are less pronounced

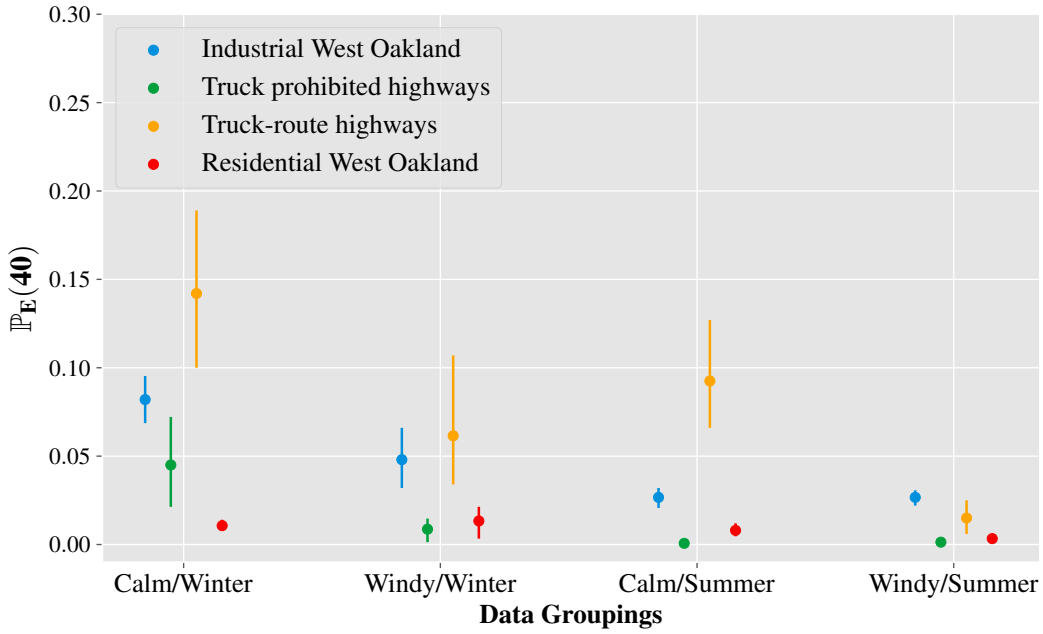

Figure 4.7: Probability of observing NO_2 concentrations above 40 ppb for groupings based on cluster, season and wind speed. Exceedance probabilities are calculated as the average of 1000 sampling simulations shown as filled circles, with vertical lines corresponding to the 25th-75th percentile ranges.

on NO_2 concentrations on highways compared to residential and industrial regions in West Oakland.

4.6.2 Exceedance probabilities

The two-step sampling process used to compute the exceedance probabilities, requires two parameters: Number of randomly selected days, N_D, and the number of samples per day, N. Here, we investigate the dependence of the calculated exceedance probabilities on these two parameters, N_D and N, respectively.

Table 4.2: Slope of linear fit to conditionally averaged NO_2 concentrations for 4 clusters during winter. Numbers in brackets refer to the p-values of the slope significance t-tests and are shown for p-values above 0.05. The boldface row corresponds to the analysis of section 4.5.2.

Δu (m/s)	Industrial West Oakland	Residential West Oakland	Truck-Prohibited Highways	Truck Route Highways
0.5	-3.16	-3.04	-0.61 (0.07)	-2.12
0.6	-3.07	-2.96	-0.51 (0.18)	-2.02
0.7	-3.00	-2.98	-0.49 (0.22)	-1.84
0.8	-3.01	-2.85	-0.49 (0.22)	-1.81
0.9	-3.42	-3.17	-0.63 (0.23)	-1.95
1.0	**-3.24**	**-3.04**	**-0.48 (0.20)**	**-1.88**
1.1	-3.17	-2.69	-0.53 (0.34)	-1.71
1.2	-3.37	-2.97	-0.38 (0.50)	-2.09
1.3	-3.03	-2.83	-0.73 (0.22)	-1.87
1.4	-3.11	-3.16	-0.54 (0.46)	-1.82 (0.11)
1.5	-2.92	-2.94	-0.44 (0.43)	-1.79 (0.07)

Sensitivity to number of randomly selected days, N_D The exceedance probabilities were calculated as described in section 4.4.2 for number of randomly selected days between 10 and 20 days. For each N_D, the average exceedance probabilities for 1000 simulations were computed for each cluster under each wind/season conditions. The resulting average exceedance probabilities showed very little dependence on N_D with all values staying within 10% of the original average exceedance probabilities plotted in Figure 4.7.

Sensitivity to number of samples per day, N Similarly exceedance probabilities were calculated with varying number of samples per day between 100 and 500 with increments of 50. There was no observable change in exceedance probabilities when number of samples per day was increased, suggesting that the original sampling of 100 samples per day was sufficiently large and therefore did not influence the exceedance probabilities.

4.7 Conclusions

An understanding of the interaction between urban form and the temporal dynamics of air pollutants is crucial for characterizing the effects of urban development and climate change on urban air quality, and especially for understanding how different settings in a given city can be subject to different health risks. In this study, a spatio-temporal framework consisting of a spatial clustering analysis and a robust statistical analysis of wind speed effects on pollutant concentrations was presented. The framework was used to study the influence of wind speed in the reduction of NO_2 concentrations in different regions of Oakland, California during different seasons. The analysis showed that wind speed is an effective tool in reducing NO_2 levels in industrial and residential regions bounded by highways during winter. However, it was found that increased vertical mixing of pollutants caused by sources other than wind speed (e.g. moving traffic, increased surface heat fluxes during summer) can lower the effectiveness of wind speed in lowering NO_2 concentrations. Furthermore, an analysis of exceedance probabilities showed that the response of the tails of the concentration distribution differs from that of the mean concentrations. These findings coupled with projections of climate and urban development can be used as predictive tools for future air quality in urban areas. For example, if reductions in wind speeds and increases in periods of stability as observed over the past few decades continue (through either climate or urban density changes), on the basis of the current level of emissions poorer air quality is expected in residential and industrial areas of Oakland during winter [121].

The application of the proposed framework to mobile measurements in Oakland has been insightful in comparing the effects of wind speed on NO_2 concen-

trations across different clusters. However, the findings presented here are particular to the measurement domain of Oakland, and generalizing the findings to other urban areas should be done with care. On the other hand, the proposed framework can be applied to other urban areas with less consistent meteorology than Oakland, to study the effects of other prominent meteorological parameters on air quality as mediated by local land use. Furthermore, the framework could be applied to study the response of other major air pollutants such as ozone (O_3) and $PM_{2.5}$ to meteorological conditions as influenced by varying urban land form. Meanwhile, the framework could be improved through the comparison of clustering solutions resulting from other well-known clustering algorithms such as DBSCAN, HDBSCAN, and hierarchical clustering, to name a few.

By utilizing the meteorological data from one station, we captured the effect of urban form in mediating the effect of regional meteorology on intra-urban air quality. We note that an improved measurement campaign could deploy meteorological stations in the measurement area (e.g. in each cluster) or integrate anemometers onto the measurement vehicle for real-time wind speed measurements [122]. In that case, an even more robust spatio-temporal analysis can be designed to study the relationship between air quality and meteorological conditions at the neighborhood scale. Furthermore, coupled meteorological and air quality measurements can also be utilized in emission source characterization, similar to efforts in characterizing methane emission sources using mobile sensors in the oil and gas industry [23].

CHAPTER 5

CONCLUSIONS

Studies presented in this dissertation are motivated by the need for effective mitigation of air pollution which can be achieved by characterizing the state and dynamics of relevant environmental variables. The main objective is to combine deterministic and statistical models to describe the connections between pollutant concentrations, emission rates, spatial domain, and meteorological conditions in a systematic and rigorous manner.

Given the wide range of spatial and temporal scales observed in various datasets and discrepancies between scales of data and deterministic models, it is unlikely that a "one-size-fits-all" methodology can be found to synergize physics-based and data-driven models. Consequently, the approaches presented in each chapter of this dissertation while different, are all examples of novel blending of deterministic and statistical methods. In particular, in chapter 2 the statistics of the LES numerical solutions are used to quantify uncertainty terms that are derived from a physical model of instantaneous plume transport. In chapter 3, physics-based knowledge is integrated into the Bayesian inference scheme through the choice of the prior distribution of the emission rate and the likelihood function. In fact, the likelihood function fuses the physical plume transport model and the observed data before updating the prior distribution to estimate the posterior distribution over the emission rate. In chapter 4, the preprocessing steps involving the choice of land-use variables prior to clustering are based on identifying the physical processes that drive NO_2 pollution. Furthermore, our findings are verified through physical understanding of the processes, relating pollutant concentrations in each cluster to regional wind speed.

Key findings from each chapter are summarized below.

In chapter 2, a theoretical framework was described to formally express the uncertainties that are inherent in emission rate quantification via gas imaging. It was shown that these uncertainties are due to the projection of three dimensional velocity and concentration fields of plumes onto two dimensional images. Simulated data from Large eddy simulations were then used to quantify the importance of each term that contributes to the total projection uncertainties. We found that the covariance error term that is related to joint spatial fluctuations of concentration and velocity fields is responsible for about 25-35% of the projection uncertainty depending on the distance from the emission source. On the other hand, we found that while the mean velocity error grows with distance from the source, its rate of increase slows down at larger distances from the source. Therefore, we observed lower total projection uncertainties further away from the source. Moreover, the investigation of the effect of time averaging and acquisition times on the total projection uncertainties highlighted that increasing the acquisition time to 5 seconds can lead to a drop of more than 50% in the total projection errors. Our findings coupled with practical findings suggest that use of multiple control volumes at varying distances from the source and longer acquisition times as allowed by operational conditions can lead to significant drops in projection uncertainties which can pave the way for robust and rapid leak quantification through gas imaging.

In chapter 3, recursive Bayesian inference method was introduced for simultaneously estimating and detecting changes in the emission rate of a point-source. This method was applied to a series of controlled release experiments and its performance was measured according to several measures. We found

that the coefficient of variation of the measured cross-plume integrated mass concentrations is a good predictor of the performance of the Bayesian inference algorithm, with lower coefficients pointing to higher success rate in detecting changes. Furthermore, the range of the measurements was a main predictor of the probability of false alarms. Although more experiments under real world conditions are required for better evaluation of the Bayesian inference algorithm, our findings show the potential of this method by highlighting its rapid change detection and low rate of false alarms compared to current methods of fault detection [13].

In chapter 4, a spatio-temporal framework was developed to characterize the effect of regional meteorology on pollutant concentrations when mediated by local land use. Particularly, the framework was used to quantify the influence of wind speed in reduction of NO_2 concentrations in spatial clusters in Oakland, California in different seasons. We found that wind speed is effective in reducing NO_2 concentrations in industrial and residential areas that are bounded by highway during winter. In addition, it was shown that NO_2 concentrations are less susceptible to wind speed in all clusters during summer and during all seasons on highways. This result is in agreement with increased vertical mixing of pollutants by other sources than wind speed, such as moving traffic on highways, and increased surface heat fluxes during summer that lead to lower effectiveness of wind in lowering NO_2 concentrations. Our findings coupled with projections of climate and urban development can be used as predictive tools for future air quality in urban areas.

APPENDIX A

APPENDIX TO CHAPTER 3

First two sections of this Appendix (based on [34]) describe the calculation of the modeled plume-weighted advection velocity, u_e^M, and the modeled normalized vertical distribution of the mass concentrations, D_z^M. These terms are required for relating cross-plume integrated mass concentrations to the emission rate through equation (3.10) and are essential to the Bayesian inference scheme introduced in chapter 3. Section A.3 is dedicated to the assessment of the effects of different experimental conditions on the performance of the changepoint detection algorithm.

A.1 Modeling the plume-weighted advection velocity

Here, we describe the determination of $u_e^M(x_m, t)$ based on the logarithmic wind profile. In a horizontally homogeneous atmospheric boundary layer, the mean streamwise velocity $\bar{u}$ follows the logarithmic profile based on the Monin-Obukhov Similarity Theory (MOST) [45]:

$$\bar{u} = \frac{u_*}{k_v}\left[\ln\left(\frac{z}{z_0}\right) - \psi\left(\frac{z}{L}\right)\right], \tag{A.1}$$

where u_* is the friction velocity, k_v is the von Karman constant (0.4), z_0 is the surface roughness (1.0 cm for short grassland [35]), $L = -\frac{u_*^3 \overline{T}}{k_v g \overline{w'T'}}$ is the Obukhov length [45], where g is the gravitational acceleration (9.81 m/s^2), $\overline{T}$ is the mean air temperature, and $\overline{w'T'}$ is the mean covariance of the instantaneous w and T.

$\psi\left(\frac{z}{L}\right)$ is a dimensionless stability correction function [36]:

$$\psi\left(\frac{z}{L}\right) = \begin{cases} -4.7z/L & L \geq 0 \\ 2\ln\left(\frac{1+\varphi}{2}\right) + 2\ln\left(\frac{1+\varphi^2}{2}\right) - 2\tan^{-1}(\varphi) + \frac{\pi}{2} & L < 0, \end{cases} \tag{A.2}$$

where $\varphi = (1 - 16z/L)^{1/4}$.

To calculate $u_e^M(x_m, t)$, we first replace $u(x_m, y, z, t)$ in equation 3.4 with $\bar{u}$, which is a function of z and t only. Next, we model the inner integral of equation 3.4 using the Lagrangian Stochastic Model (LSM) as $D_z^M(x_m, z, t)$, as detailed in section A.2. With these modifications, $u_e^M(x_m, t)$ can be computed as follows

$$u_e^M = \int_{z_{min}}^{z_{max}} D_z^M(x_m, z, t)\bar{u}dz. \tag{A.3}$$

A.2 Modeling the normalized distribution of concentrations

In this section, we use the LSM to estimate $D_z^M(x_m, z, t)$. The LSM describes the plume dispersion by calculating the trajectories of marked fluid particles in 2D (longitudinal and vertical directions). The particle positions (x_p and z_p) in the downwind and vertical directions are calculated as

$$dx_p = \left(u_p + \bar{u}\right)dt, \tag{A.4}$$

$$dz_p = w_p dt, \tag{A.5}$$

where dt is the time step. u_p and w_p are the particle's Lagrangian velocity in the x and z directions, which follows the generalized Langevin equation [75]:

$$du_p = a_u dt + b_u dW, \tag{A.6}$$

$$dw_p = a_w dt + b_w dW, \tag{A.7}$$

where dW is an incremental Wiener process with a zero mean and a variance of dt. a and b are parameters that need to satisfy Kolmogorov's hypothesis of local isotropy in a Lagrangian frame of reference [123] and the well-mixed condition [75]. According to the simplest solution and considering a flat homogeneous surface layer, a and b can be formulated as:

$$a_u = -\frac{b_u^2}{A}\left(\sigma_w^2 u_p - \overline{u'w'}w_p\right) + \frac{1}{A}\left(\sigma_w^2 \frac{\partial \sigma_u^2}{\partial z} u_p w_p - \overline{u'w'}\frac{\partial \sigma_u^2}{\partial z} w_p^2\right), \tag{A.8}$$

$$a_w = -\frac{b_w^2}{A}\left(\sigma_u^2 w_p - \overline{u'w'}u_p\right) + \frac{1}{A}\left(-\overline{u'w'}\frac{\partial \sigma_w^2}{\partial z} u_p w_p + \sigma_u^2 \frac{\partial \sigma_w^2}{\partial z} w_p^2\right) + \frac{1}{2}\frac{\partial \sigma_w^2}{\partial z}, \tag{A.9}$$

$$b_u = \sigma_w \sqrt{2/T_L}, \tag{A.10}$$

$$b_w = \sigma_w \sqrt{2/T_L}, \tag{A.11}$$

where $A = 2\left(\sigma_u^2 \sigma_w^2 - \overline{u'w'}^2\right)$, σ_u and σ_w are the standard deviations of the Eulerian velocity in the longitudinal and vertical directions, $\overline{u'w'}$ is the Reynolds stress, and T_L is the Lagrangian velocity time scale.

To solve Equations (A.4) - (A.11), vertical profiles of $\bar{u}, \sigma_u, \sigma_w, \overline{u'w'}$ and T_L are required. The Reynolds stress $\overline{u'w'} = -u_*^2$ is assumed to be a constant in the surface layer [36]. Since measurements were available only at a single height during each experiment, the MOST was applied to describe the vertical profiles of all wind statistics required by the LSM (i.e., $\bar{u}, \sigma_u$ and σ_w). More specifically, $\bar{u}$ is described by logarithmic wind profile of equation (A.1). Meanwhile, σ_u and σ_w can be described as [124]:

$$\sigma_u = 3.7 \times u_* \left(1 - 3\frac{z}{L}\right)^{1/3}, \ \sigma_w = 1.26 \times u_* \left(1 - 3\frac{z}{L}\right)^{1/3}. \tag{A.12}$$

The empirical parameters 3.7 and 1.26 are estimated by fitting the measured σ_u and σ_w by the 3D sonic sonic anemometer against $(1 - 3z/L)^{1/3}$ [34]. In order to solve the Langevin equation, we also estimate the Lagrangian time scale. T_L

can be estimated using K-theory as follows [125]

$$T_L = \frac{K}{\sigma_w^2}, \tag{A.13}$$

where K is the diffusion coefficient [126]

$$K = \frac{k_v u_* z}{\psi_h}, \tag{A.14}$$

and ψ_h is the stability correction [127]

$$\psi_h = \begin{cases} \left(1 - 3\frac{z}{L}\right) & z/L \geq 0 \\ 0.32\left(0.037 - \frac{z}{L}\right)^{-1/3} & z/L < 0. \end{cases} \tag{A.15}$$

During an LSM run, particles' velocity and location are stored in the $x - z$ domain. Given a point source with a unit emission rate, the predicted mean concentration after integrating over y, denoted as C^y_{LSM}, can be described as [128, 129]:

$$C^y_{LSM}(x, z, t) = \frac{1}{N_p \Delta x} \sum \frac{1}{|w_p(x, z)|}, \tag{A.16}$$

where N_p is the number of fluid particles, Δx and Δz are the grid spacing of the LSM in x and z directions, respectively. The LSM modeled $D^M_z(x_m, z, t)$ can be computed as:

$$D^M_z(x_m, z, t) = \frac{C^y_{LSM}(x_m, z, t)}{\sum_{z_{min}}^{z_{max}} C^y_{LSM}(x_m, z, t)\Delta z}, \tag{A.17}$$

where x_{max} and z_{max} are used to denote the size of the computational domain in the x and z directions, respectively.

A computational domain of (x_m, z_m) = (100m, 20m) is used with the grid cell size of $(\Delta x, \Delta z$ = (0.5m, 0.01m) and dt is determined dynamically as dt = $\min[0.02T_L, \Delta z/w(t - dt)]$ to satisfy the necessary condition of $dt \ll T_L$ to prevent a large jump in the vertical direction [125]. For each run, a total of $N_p = 10^6$ particles are released from the source and their trajectories are calculated using

Table A.1: Summary of parameters used in LSM.

Name	**Value**
Number of fluid particles (N_p)	10^6
Computational domain size (x_{max}, z_{max})	100, 20 (m)
Computational grid size ($\Delta x, \Delta z$)	0.5, 0.01 (m)
Surface roughness (z_0)	0.01 (m)
von Karman constant (k_v)	0.4
Height of sensor (z_m)	1.3 (m)
Height of source (z_s)	0.02 (m)
Friction velocity (u_*)	case dependent
Standard deviation of u and w (σ_u, σ_w)	case dependent
Obukhov length (L)	case dependent

the LSM. The ground and the boundary-layer top (set to be 1 km for neutral and unstable conditions) are considered as perfect reflectors, such that a particle will be perfectly reflected back in the vertical direction and the sign of both u_p and w_p reversed [125]. Particles can only exit at the end of the domain when $x(t) > x_{max}$. A summary of parameters used in the LSM is shown in Table A.1.

Calculating $D_z^M(x_m, z, t)$ at x_m = 20 and 30m without considering the obstacle may be acceptable, however, this assumption will introduce large errors at x_m =10m [34]. For simplicity, we assume that $D_z^M(x_m, z, t)$ is vertically well-mixed by obstacle-injected wake eddies from ground to $1.5 \times z_F$ at x_m = 10m, where z_F is the height of the obstacle. Therefore, we write

$$D_z^M(10, z < 1.5z_F, t) = \frac{1}{1.5z_F} \sum_{0}^{1.5z_F} D_z^M(10, z, t)\Delta z. \tag{A.18}$$

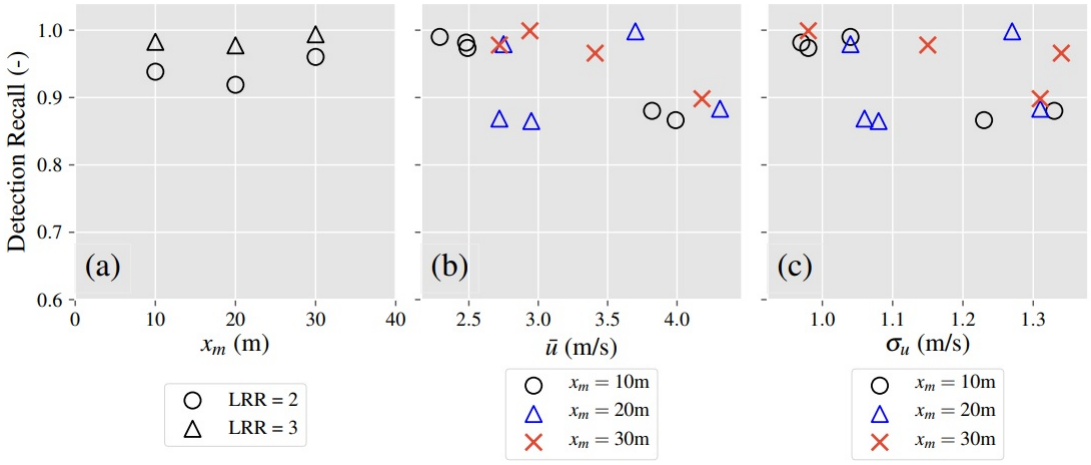

Figure A.1: Assessment of the effect of (a) source-to-sensor distance (x_m), (b) mean wind speed ($\bar{u}$), and (c) standard deviation of wind speed (σ_u) on the performance of the changepoint detection algorithm represented by detection recall.

A.3 Assessment of the effects of source-to-sensor distance and wind speed on changepoint detection performance

Here, we examine the effects of different experimental conditions, namely the downwind distance (x_m) and wind speed statistics ($\bar{u}, \sigma_u$) on the performance of the changepoint detection algorithm of chapter 3. For these assessments, the leak rate ratio is set to a value of 2 (also 3 in Figure A.1, with similar results observed for other investigated ratios.

To highlight the effect of downwind distance, we averaged the detection recall over experiments sharing the same x_m. As shown in Figure A.1a, we observe little dependence of changepoint detection performance on x_m with no obvious trend being established between the two variables. This result is expected due to the little dependence of the leak estimation through Bayesian inference on x_m when obstacles are present [34]. Similarly, Figures A.1b and A.1c show little cor-

relation between the detection recall and wind speed statistics. These findings highlight the robustness of the changepoint detection algorithm under varying wind conditions.

APPENDIX B

APPENDIX TO CHAPTER 4

In this Appendix, we first provide the details of the snapping procedure used to move the raw location of each 1-Hz mobile measurement to the nearest 30-m road segment. In section B.2, the steps followed to gather and calculate all the geographic covariates used in the clustering analysis are described in detail. The analysis of the fixed site monitor that motivated the stratification of data by season is presented in section B.3. Finally, the data pre-processing approach involving Principal Component Analysis (PCA) of the geographic covariates is described.

B.1 Spatial coordinate snapping

Each 1-Hz measurement of NO_2 concentration corresponds to a GPS coordinate pair (longitude and latitude in the WGS-84 geographic coordinate system. With each street traversed multiple times throughout the course of the mobile campaign, small spatial variation in the recorded GPS coordinates is inevitable. This variation is due to a combination of vehicle motion and measurement uncertainty of the GPS unit (nominally ±3m for clear-sky conditions and degraded by the presence of buildings and trees). To analyze the statistics of NO_2 measurements at the same location, it is helpful to aggregate nearby observations to a consistent set of pre-determined locations. This spatial "snapping" is achieved using the following approach.

First, the road line geometry shapefile for Oakland is obtained from Open-

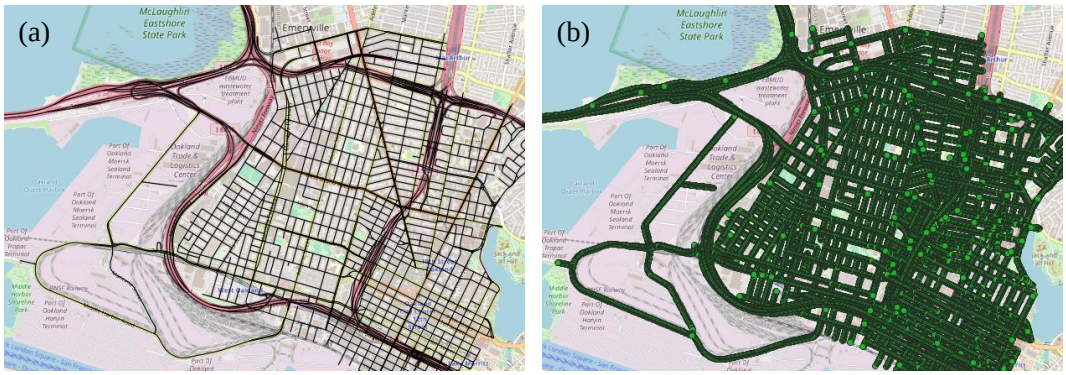

Figure B.1: (a) The roadline geometry in QGIS and its conversion to (b) point geometry with 30-meter spacing. Map data by OSM.

StreetMap (OSM) and converted to point geometry at 30m spacing using a built-in function of QGIS (an open source Geographic Information System). Figure B.1 shows the conversion from line to point geometry for West and Downtown Oakland in QGIS. Each point in the converted geometry corresponds to the midpoint of an individual 30-meter road segment. Then, for each raw GPS measurement, a nearest-neighbor algorithm (Python SciPy "ckdnearest" algorithm) was used to locate the nearest midpoint of a 30-meter road segment as available in the point geometry shapefile [107].

B.2 Calculation of geographic covariates

For each 30-meter road segment in the point geometry shapefile, a total of 26 binary and continuous geographic covariates were calculated as described in the following sections with a summary of the covariates presented in Table B.1.

B.2.1 Road type classifications and road length

The road line geometry of OSM groups the roads in Oakland into 14 separate categories. Following Apte et al., we create binary variables classifying road types into 3 distinct categories of Highways, Major arterials, and Residential roads [88]. Our road type classifications correspond to OSM groups as follows:

- **Highways**: Motorway, motorway link, trunk, and trunk link
- **Major arterials**: Primary, primary link, secondary, secondary link, service, tertiary, tertiary link, and unclassified
- **Residential**: Living street, and residential,

with Figure B.2 presenting the road type classification of 30-meter road segments for West and Downtown Oakland in QGIS.

Moreover, the OSM data is used to calculate the total road lengths for highways, major arterials, residential roads, and total roads within 50m circular buffers around each road segment.

Further, city of Oakland classifies certain routes as designated heavy-duty truck routes for which we create a binary classification variable highlighting whether a road segment is on a designated heavy-duty truck route. Similarly, a binary classification variable is created based on whether a road segment is on a road where heavy-duty trucks are explicitly prohibited by the city of Oakland.

Figure B.2: Classification of road segments by road type with highways, major arterials, and residential roads represented by yellow, red, and blue circles, respectively. Map data by OSM.

B.2.2 City of Oakland zoning

The zoning map published by the city of Oakland is used to create binary variables representing commercial, industrial and residential zoning [130]. Figure B.3 presents the correspondence of 30-meter road segments to each of the 3 zoning categories for West and Downtown Oakland in QGIS.

B.2.3 Distance to point sources

For each 30-meter road segment we calculate the distance to potential air pollution sources including railway stations, airports, ports, National Priority Listing (NPL) sites, and Toxic Release Inventory (TRI) sites. The information related to each pollution source was gathered as follows:

- **Railway stations**: The locations of stations were obtained through OSM.
- **Airports**: The locations of major airports close to the study area (SFO and

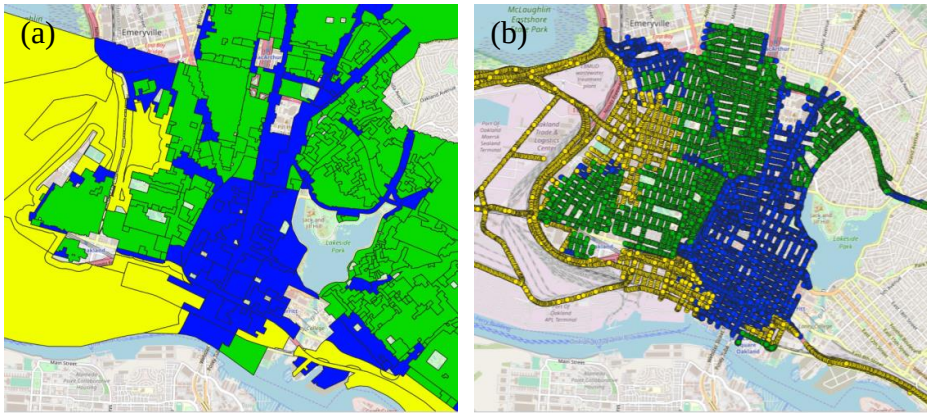

Figure B.3: (a) Zoning polygons used to classify (b) zoning of road segments with residential, commercial and industrial zones represented by green, blue and yellow circles, respectively. Map data by OSM.

OAK) were obtained through OSM.

- **Ports**: Locations of ports and port facilities were obtained from the U.S. Army Corps of Engineers data on principal port locations and their associated facilities. Data was downloaded from the State of California Geoportal (`https://gis.data.ca.gov/datasets/`; accessed October 15, 2020).
- **NPL**: U.S. EPA superfund NPL sites in Alameda county were downloaded from the U.S. EPA website [131].
- **TRI**: U.S. EPA TRI sites were downloaded from the U.S. EPA website [132].

B.2.4 Mean elevation, population density and NDVI

The mean elevation, population density and Normalized Difference Vegetative Index (NDVI) for each 30-meter road segment are calculated within 50m circular

buffers. The population density is calculated using the 2010 census tract population data as follows. First, a 30-meter grid raster file is created from the census tract population shapefile polygon under the assumption of uniform population density within each census tract. The total population in the census tract is then distributed evenly over each raster grid cell. Finally, the population density for each road segment is calculated based on the raster file and the area of the circular buffer.

The mean elevation is calculated using the Shuttle Radar Topography Mission (SRTM) digital elevation data obtained through Google Earth Engine (GEE) (`https://earthengine.google.com/`; accessed October 1, 2020) [133]. To download data using GEE, the point geometry of midpoints of 30-meter road segments are exported from QGIS and imported into GEE. The code snippet used in GEE to gather elevation data within 50m circular buffers is illustrated in Figure B.4 with "pfg" denoting the imported road segment point geometry.

NDVI is calculated using the Landsat8 data archived at the U.S. Geological Survey (USGS) Earth Resources Observation and Science (EROS) through GEE [134]. The code snippet used in GEE to gather NDVI data within 50m circular buffers is illustrated in Figure B.5.

B.2.5 National land cover database (NLCD)

To assign a numeric value to the land cover surrounding 30-meter road segments, we assign explanatory variables to each land cover type. For a land cover of type l, for each road segment, the explanatory variable corresponding

```
Imports (2 entries)
  var srtm: Image "NASA SRTM Digital Elevation 30m" (1 band)
  var pfg: Table users/[redacted]/points_for_gee_shp
// Create the buffering functions
var bufferPoly50 = function(feature) {
  return feature.buffer(50);
};

// define function for calculating mean elevation for the buffer
var meanDic = function(func){
  var f50 = bufferPoly50(func);
  var xcoord = func.get('Lon');
  var ycoord = func.get('Lat');
  var temp50 = srtm.reduceRegion({
    reducer: ee.Reducer.mean(),
    geometry: f50.geometry(),
    scale: 30
  });
  return ee.Feature(func.geometry(),{'Lon': xcoord,'Lat':ycoord ,
  'elevation50':temp50.get('elevation')});
};

// apply the mean elevation function to every point
var meanDict = pfg.map(meanDic);

// export the created table to google drive
Export.table.toDrive({
  collection: meanDict,
  description:'pfg_elevation',
  fileFormat: 'CSV'
});
```

Figure B.4: GEE code snippet for calculating and downloading mean elevation in 50m circular buffers around road segments.

to l is calculated as [104]:

$$LC^{(l)} = \frac{1}{n}\sum_{i=1}^{n} I_i^{(l)} \tag{B.1}$$

where $LC^{(l)}$ is the percent of land cover type l within a radius of 50m of the road segment, $I_i^{(l)}$ is an indicator variable equal to 1 if the i'th pixel surrounding the road segment is of type l, and zero otherwise, and n is the number of pixels within the 50m circular buffer around the road segment. We create variables for NLCD land cover types of Barren Land, Developed Open, Developed Low, Developed Medium, Developed High, and Cultivated Crops.

The NLCD image data is obtained from the USGS NLCD using GEE. The

```
Imports (3 entries)
  var l8: ImageCollection "USGS Landsat 8 Collection 1 Tier 1 TOA Reflectance" (12 bands)
  var pfg: Table users/[redacted]/points_for_gee_shp
  var roi: Point (-122.28, 37.82)
// function for calculating NDVI from image bands
function addNDVI(image) {
  var ndvi = image.normalizedDifference(['B5', 'B4']);
  return image.addBands(ndvi);
  }

// Filter around Oakland during relevant dates
var filtered = l8.filterDate('2015-06-15', '2017-07-15')
.filterBounds(roi);
// add ndvi to image collection
var with_ndvi = ee.ImageCollection(filtered.map(addNDVI));

// find the median ndvi for study period
var image = with_ndvi.median().select('nd');

// Create the buffering functions
var bufferPoly50 = function(feature) {
  return feature.buffer(50);
};

// define function for calculating mean NDVI for circular buffer
var meanDic = function(func){
  var f50 = bufferPoly50(func);
  var xcoord = func.get('Lon');
  var ycoord = func.get('Lat');
  var temp50 = image.reduceRegion({
    reducer: ee.Reducer.mean(),
    geometry: f50.geometry(),
    scale: 30
  });
  return ee.Feature(func.geometry(),{'Lon': xcoord,'Lat':ycoord ,
  'ndvi50':temp50.get('nd')});
};

// apply the average elevation function to every point
var meanDict = pfg.map(meanDic);
// export the created table to google drive
Export.table.toDrive({
  collection: meanDict,
  description:'pfg_NDVI',
  fileFormat: 'CSV'
});
```

Figure B.5: GEE code snippet for calculating and downloading mean NDVI in 50m circular buffers around road segments.

code snippet used in GEE to gather the NLCD image data is presented in Figure B.6. This data was then used to calculate the explanatory variables described above.

```
Imports (3 entries)
  var nlcd: ImageCollection "NLCD: USGS National Land Cover Database"
  var roi: Point (-122.28, 37.82)
  var pfg: Table users/[redacted]/points_for_gee_shp
// Filter around Oakland during relevant dates
var filtered = nlcd.filterDate('2015-06-15', '2017-07-15')
.filterBounds(roi);
var lc = filtered.first().select('landcover');
// Create the buffering functions
var bufferPoly50 = function(feature) {
  return feature.buffer(50);
};
// define function for calculating NLCD histogram for circular buffer
var meanDic = function(func){
  var f50 = bufferPoly50(func);
  var xcoord = func.get('Lon');
  var ycoord = func.get('Lat');
  var temp50 = lc.reduceRegion({
    reducer: ee.Reducer.frequencyHistogram(),
    geometry: f50.geometry(),
    scale: 30
  });
  return ee.Feature(func.geometry(),{'Lon':xcoord,'Lat':ycoord,
  'nlcd50':temp50.get('landcover')});
};
// apply the NLCD histogram function to every point
var meanDict = pfg.map(meanDic);

// export the created table to google drive
Export.table.toDrive({
  collection: meanDict,
  description:'pfg_NLCD',
  fileFormat: 'CSV'
});
```

Figure B.6: GEE code snippet for calculating and downloading frequency histogram of NLCD variables in 50m circular buffers around road segments.

B.3 Analysis of fixed site monitor in West Oakland

To investigate the effect of seasonal changes on NO_2 concentrations in Oakland, we used data from a fixed-site regulatory monitor located at the Oakland West site managed by the Bay Area Air Quality Management District (BAAQMD). This site is located in a mixed commercial-industrial with a distance of 30m from a local arterial. Instruments at this site include a chemiluminescence NO_2

Table B.1: List of calculated land-use (geographic) covariates used in the cluster analysis.

Variable name	**Description**
Highway road type	Binary road type variable
Major arterial road type	Binary road type variable
Residential road type	Binary road type variable
Highway length	Length of highway road type within a 50m circular buffer
Major length	Length of major road type within a 50m circular buffer
Residential length	Length of residential road type within a 50m circular buffer
Total length	Length of all roads within a 50m circular buffer
Truck-route	Binary road type variable in addition to the 3 categories
Truck-prohibited route	Binary road type variable in addition to the 3 categories
Residential zone	Binary variable related to city of Oakland zoning
Commercial zone	Binary variable related to city of Oakland zoning
Industrial zone	Binary variable related to city of Oakland zoning
Distance to port/port facilities	Minimum distance to point sources
Distance to airport	Minimum distance to point sources
Distance to railway stations	Minimum distance to point sources
Distance to TRI	Minimum distance USEPA Toxic Release Inventory
Distance to NPL	Minimum distance USEPA National Priority Listing
Elevation	Mean elevation within a 50m circular buffer
Population	Mean population density within a 50m circular buffer
NDVI	Average Normalized Difference Vegetative Index within a 50m circular buffer
Land-cover: Barren Land Land-cover: Developed Open Land-cover: Developed Low Land-cover: Developed Medium Land-cover: Developed High Land-cover: Cultivated Crops	Explanatory variables created for each land cover variable based on the National Land Cover Database (NLCD). Each explanatory variable corresponds to the percent of land cover of each type calculated within a 50m circular buffer.

monitor (Model Teco 42i, Thermo-Fisher Scientific) [88].

We collected hourly daytime (7am-7pm) NO_2 concentration data for the 7 year period of 2011-2017 to analyze monthly changes in NO_2 concentrations at West Oakland site. We then calculated a daily average concentration for each day in the 7 year period, and used these daily averages to calculate conditionally averaged concentrations based on the month of the year. The daily averages

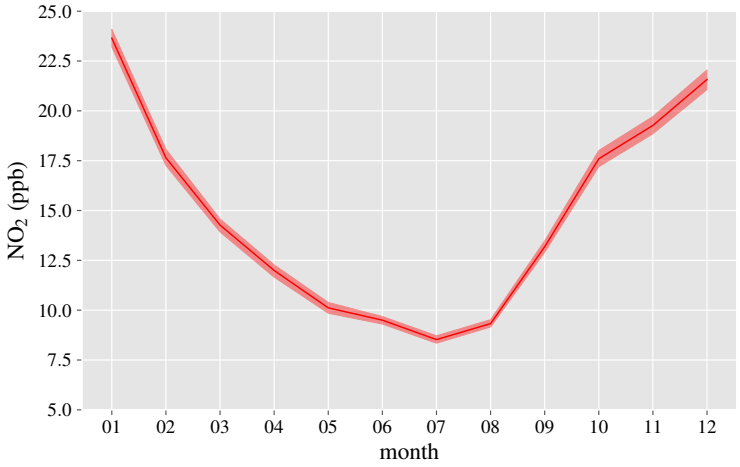

Figure B.7: Daily average of daytime (7am-7pm) NO_2 concentrations averaged by month of year as measured by the BAAQMD Oakland West fixed site monitor with the shaded region corresponding to the 95% confidence interval.

were then used in a bootstrapping significance test analysis to establish 95% confidence intervals for the computed conditionally averaged concentrations [80] with the results depicted in Figure B.7.

The results show that the NO_2 concentrations during summer are significantly lower than winter with a 63% drop from January to July. This finding motivates the seasonal stratification of the mobile measurement data described in section 4.3.

B.4 Principal component analysis of land-use data

In addition to feature reduction, PCA is useful in investigating the linear correlation between variables. Here, we use PCA to identify how land-use variables are related with respect to the two primary axes of data. As shown in Figure B.8, the first two principal components capture 43% of the total variation in the

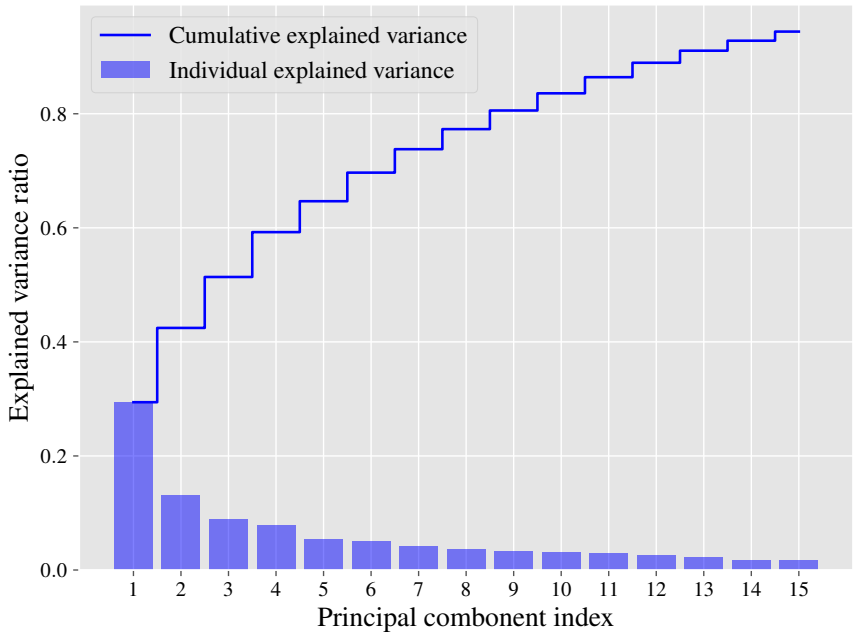

Figure B.8: Ratio of explained variance by the first 15 principal components.

dataset, with the first and second components explaining 29 and 14 percent of the total variance, respectively. Because these two components describe a substantial share of the variance in the land-use data, a two-dimensional projection of the data onto them (Figure B.9) allows for visualizing patterns of similarity among the covariates [135].

Figure B.9 indicates that the road length and road type variables are strongly correlated as expected, since a road segment belonging to a certain road type is likely to be surrounded by road segments of the same type. In addition, location, distance and zoning variables are also highly correlated suggesting a lot of redundancy in the land-use features. We hypothesize that the first principal component relates primarily to road type variables, while the second principal component is mainly influenced by the coordinates of the road segment (e.g., segments in the North West are the farthest from segments in the South East).

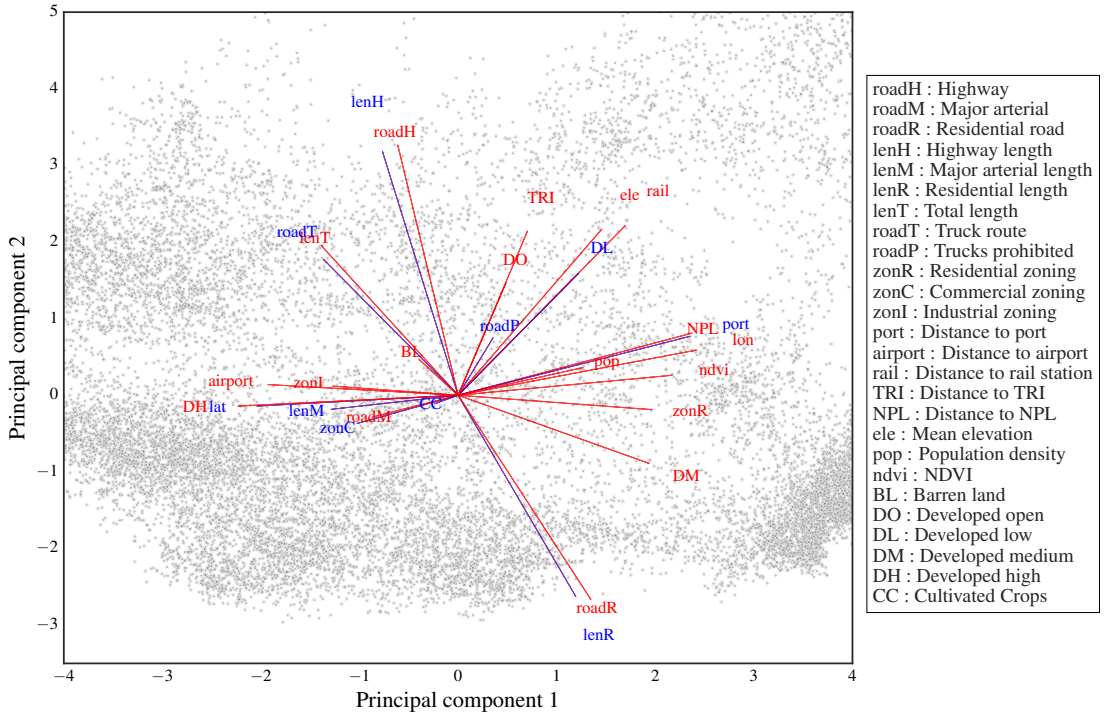

Figure B.9: Labeled biplot diagram of all land-use covariates onto first and second principal components.

BIBLIOGRAPHY

[1] D. J. Jacob and D. A. Winner, "Effect of climate change on air quality," *Atmospheric Environment*, vol. 43, pp. 51–63, Jan. 2009.

[2] X. Zhou, S. Yoon, S. Mara, M. Falk, T. Kuwayama, T. Tran, L. Cheadle, J. Nyarady, B. Croes, E. Scheehle, J. D. Herner, and A. Vijayan, "Mobile sampling of methane emissions from natural gas well pads in California," *Atmospheric Environment*, vol. 244, p. 117930, Jan. 2021.

[3] D. Zavala-Araiza, R. A. Alvarez, D. R. Lyon, D. T. Allen, A. J. Marchese, D. J. Zimmerle, and S. P. Hamburg, "Super-emitters in natural gas infrastructure are caused by abnormal process conditions," *Nature Communications*, vol. 8, p. 14012, Jan. 2017.

[4] R. Rai and C. K. Sahu, "Driven by Data or Derived Through Physics? A Review of Hybrid Physics Guided Machine Learning Techniques With Cyber-Physical System (CPS) Focus," *IEEE Access*, vol. 8, pp. 71050–71073, 2020. Conference Name: IEEE Access.

[5] A. Karpatne, G. Atluri, J. H. Faghmous, M. Steinbach, A. Banerjee, A. Ganguly, S. Shekhar, N. Samatova, and V. Kumar, "Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, pp. 2318–2331, Oct. 2017. Conference Name: IEEE Transactions on Knowledge and Data Engineering.

[6] S. G. Perry, D. K. Heist, L. H. Brouwer, E. M. Monbureau, and L. A. Brixey, "Characterization of pollutant dispersion near elongated buildings based on wind tunnel simulations," *Atmospheric Environment*, vol. 142, pp. 286–295, Oct. 2016.

[7] R. L. Petersen, S. A. Guerra, and A. S. Bova, "Critical review of the building downwash algorithms in AERMOD," *Journal of the Air & Waste Management Association*, vol. 67, pp. 826–835, Aug. 2017. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10962247.2017.1279088.

[8] C. S. Cheng, M. Campbell, Q. Li, G. Li, H. Auld, N. Day, D. Pengelly, S. Gingrich, and D. Yap, "A Synoptic Climatological Approach to Assess Climatic Impact on Air Quality in South-central Canada. Part II: Future Estimates," *Water, Air, and Soil Pollution*, vol. 182, pp. 117–130, June 2007.

[9] C. Y. C. Lin, D. J. Jacob, and A. M. Fiore, "Trends in exceedances of the ozone air quality standard in the continental United States, 1980â€"1998," *Atmospheric Environment*, vol. 35, pp. 3217–3228, July 2001.

[10] ARPA-E, "Methane Observation Networks with Innovative Technology to Obtain Reductions." Available online: https://arpa-e.energy.gov/technologies/programs/monitor, 2015. (Accessed on 30 June 2021).

[11] R. A. Alvarez, D. Zavala-Araiza, D. R. Lyon, D. T. Allen, Z. R. Barkley, A. R. Brandt, K. J. Davis, S. C. Herndon, D. J. Jacob, A. Karion, E. A. Kort, B. K. Lamb, T. Lauvaux, J. D. Maasakkers, A. J. Marchese, M. Omara, S. W. Pacala, J. Peischl, A. L. Robinson, P. B. Shepson, C. Sweeney, A. Townsend-Small, S. C. Wofsy, and S. P. Hamburg, "Assessment of methane emissions from the U.S. oil and gas supply chain," *Science*, vol. 361, pp. 186–188, July 2018.

[12] A. P. Ravikumar and A. R. Brandt, "Designing better methane mitigation policies: the challenge of distributed small sources in the natural gas sector," *Environmental Research Letters*, vol. 12, p. 044023, Apr. 2017.

[13] Scientific Aviation, "Systematic Observations of Facility Intermittent Emissions (SOOFIE)." Available online: https://www.scientificaviation.com/soofie/, 2021. (Accessed on 15 August 2021).

[14] Q. Wang, X. Chen, A. N. Jha, and H. Rogers, "Natural gas from shale formation – The evolution, evidences and challenges of shale gas revolution in United States," *Renewable and Sustainable Energy Reviews*, vol. 30, pp. 1–28, Feb. 2014.

[15] U.S. Energy Information Administration (EIA), "Today in energy." Available online: https://www.eia.gov/todayinenergy/detail.php?id=43115, 2020. (Accessed on 19 February 2021).

[16] R. A. Alvarez, S. W. Pacala, J. J. Winebrake, W. L. Chameides, and S. P. Hamburg, "Greater focus needed on methane leakage from natural gas infrastructure," *Proceedings of the National Academy of Sciences*, vol. 109, pp. 6435–6440, Apr. 2012.

[17] U.S. Environmental Protection Agency, "New source performance standards; oil and natural gas sector: emission standards for new, recon-

structed, and modified sources," *Federal Register*, vol. 81, pp. 35824 – 35942, 2016.

[18] D. R. Johnson, A. N. Covington, and N. N. Clark, "Methane Emissions from Leak and Loss Audits of Natural Gas Compressor Stations and Storage Facilities," *Environmental Science & Technology*, vol. 49, pp. 8132–8138, July 2015.

[19] T. K. Flesch, L. A. Harper, R. L. Desjardins, Z. Gao, and B. P. Crenna, "Multi-Source Emission Determination Using an Inverse-Dispersion Technique," *Boundary-Layer Meteorology*, vol. 132, pp. 11–30, July 2009.

[20] T. Krings, K. Gerilowski, M. Buchwitz, M. Reuter, A. Tretner, J. Erzinger, D. Heinze, U. Pflüger, J. P. Burrows, and H. Bovensmann, "MAMAP – a new spectrometer system for column-averaged methane and carbon dioxide observations from aircraft: retrieval algorithm and first inversions for point source emission rates," *Atmospheric Measurement Techniques*, vol. 4, pp. 1735–1758, Sept. 2011.

[21] D. M. Tratt, K. N. Buckland, J. L. Hall, P. D. Johnson, E. R. Keim, I. Leifer, K. Westberg, and S. J. Young, "Airborne visualization and quantification of discrete methane sources in the environment," *Remote Sensing of Environment*, vol. 154, pp. 74–88, Nov. 2014.

[22] R. M. Duren, A. K. Thorpe, K. T. Foster, T. Rafiq, F. M. Hopkins, V. Yadav, B. D. Bue, D. R. Thompson, S. Conley, N. K. Colombi, C. Frankenberg, I. B. McCubbin, M. L. Eastwood, M. Falk, J. D. Herner, B. E. Croes, R. O. Green, and C. E. Miller, "California's methane super-emitters," *Nature*, vol. 575, pp. 180–184, Nov. 2019.

[23] J. D. Albertson, T. Harvey, G. Foderaro, P. Zhu, X. Zhou, S. Ferrari, M. S. Amin, M. Modrak, H. Brantley, and E. D. Thoma, "A Mobile Sensing Approach for Regional Surveillance of Fugitive Methane Emissions in Oil and Gas Production," *Environmental Science & Technology*, vol. 50, pp. 2487–2497, Mar. 2016.

[24] X. Zhou, X. Peng, A. Montazeri, L. E. McHale, S. GaÃŸner, D. R. Lyon, A. P. Yalin, and J. D. Albertson, "Mobile Measurement System for the Rapid and Cost-Effective Surveillance of Methane and Volatile Organic Compound Emissions from Oil and Gas Production Sites," *Environmental Science & Technology*, vol. 55, pp. 581–592, Jan. 2021. Publisher: American Chemical Society.

[25] X. Zhou, F. H. Passow, J. Rudek, J. C. von Fisher, S. P. Hamburg, and J. D. Albertson, "Estimation of methane emissions from the U.S. ammonia fertilizer industry using a mobile sensing approach," *Elementa: Science of the Anthropocene*, vol. 7, May 2019.

[26] J. Sandsten and M. Andersson, "Volume flow calculations on gas leaks imaged with infrared gas-correlation," *Optics Express*, vol. 20, pp. 20318–20329, Aug. 2012.

[27] M. Galfalk, G. Olofsson, P. Crill, and D. Bastviken, "Making methane visible," *Nature Climate Change*, vol. 6, pp. 426–430, Apr. 2016.

[28] N. Hagen, R. T. Kester, and C. Walker, "Real-time quantitative hydrocarbon gas imaging with the gas cloud imager (GCI)," in *Chemical, Biological, Radiological, Nuclear, and Explosives (CBRNE) Sensing XIII*, vol. 8358, p. 83581J, International Society for Optics and Photonics, May 2012.

[29] L. C. Gui and W. Merzkirch, "A method of tracking ensembles of particle images," *Experiments in Fluids*, vol. 21, pp. 465–468, Nov. 1996.

[30] M. Galfalk and D. Bastviken, "Remote sensing of methane and nitrous oxide fluxes from waste incineration," *Waste Management*, vol. 75, pp. 319–326, May 2018.

[31] M. Galfalk, G. Olofsson, and D. Bastviken, "Approaches for hyperspectral remote flux quantification and visualization of GHGs in the environment," *Remote Sensing of Environment*, vol. 191, pp. 81–94, Mar. 2017.

[32] A. P. Ravikumar, J. Wang, and A. R. Brandt, "Are Optical Gas Imaging Technologies Effective For Methane Leak Detection?," *Environmental Science & Technology*, vol. 51, pp. 718–724, Jan. 2017.

[33] A. P. Ravikumar, J. Wang, M. McGuire, C. S. Bell, D. Zimmerle, and A. R. Brandt, ""Good versus Good Enough?" Empirical Tests of Methane Leak Detection Sensitivity of a Commercial Infrared Camera," *Environmental Science & Technology*, vol. 52, pp. 2368–2374, Feb. 2018.

[34] X. Zhou, A. Montazeri, and J. D. Albertson, "Mobile sensing of point-source gas emissions using Bayesian inference: An empirical examination of the likelihood function," *Atmospheric Environment*, vol. 218, p. 116981, Dec. 2019.

[35] H. A. Panofsky and J. A. Dutton, *Atmospheric Turbulence: Models and Methods for Engineering Applications*. Hoboken, NJ, USA: John Wiley & Sons., first ed., 1983.

[36] R. B. Stull, *An Introduction to Boundary Layer Meteorology*. Dordrecht, Netherlands: Kluwer Academic Publishers, first ed., 1988.

[37] J. D. Albertson, *Large eddy simulation of land-atmosphere interaction*. University of California, Davis, 1996.

[38] E. Bou-Zeid, C. Meneveau, and M. Parlange, "A scale-dependent Lagrangian dynamic model for large eddy simulation of complex turbulent flows," *Physics of Fluids*, vol. 17, p. 025105, Jan. 2005.

[39] Y.-H. Tseng, C. Meneveau, and M. B. Parlange, "Modeling Flow around Bluff Bodies and Predicting Urban Dispersion Using Large Eddy Simulation," *Environmental Science & Technology*, vol. 40, pp. 2653–2662, Apr. 2006.

[40] Q. Li, E. Bou-Zeid, W. Anderson, S. Grimmond, and M. Hultmark, "Quality and reliability of LES of convective scalar transfer at high Reynolds numbers," *International Journal of Heat and Mass Transfer*, vol. 102, pp. 959–970, Nov. 2016.

[41] D. R. Caulton, Q. Li, E. Bou-Zeid, J. P. Fitts, L. M. Golston, D. Pan, J. Lu, H. M. Lane, B. Buchholz, X. Guo, J. McSpiritt, L. Wendt, and M. A. Zondlo, "Quantifying uncertainties from mobile-laboratory-derived emissions of well pads using inverse Gaussian methods," *Atmospheric Chemistry and Physics*, vol. 18, pp. 15145–15168, Oct. 2018.

[42] F. T. M. Nieuwstadt and J. P. J. M. M. de Valk, "A large eddy simulation of buoyant and non-buoyant plume dispersion in the atmospheric boundary layer," *Atmospheric Environment (1967)*, vol. 21, pp. 2573–2587, Jan. 1987.

[43] P. J. Mason, "Large-eddy simulation of dispersion in convective boundary layers with wind shear," *Atmospheric Environment. Part A. General Topics*, vol. 26, pp. 1561–1571, June 1992.

[44] J. C. Weil, P. P. Sullivan, and C.-H. Moeng, "The Use of Large-Eddy Simulations in Lagrangian Particle Dispersion Models," *Journal of the Atmospheric Sciences*, vol. 61, pp. 2877–2887, Dec. 2004.

[45] J. C. Kaimal and J. J. Finnigan, *Atmospheric Boundary Layer Flows: Their Structure and Measurement*. Oxford, UK: Oxford Universtiy Press, 1994.

[46] T. F. Stocker, *Climate Change 2013: The Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, 2014. Google-Books-ID: o4gaBQAAQBAJ.

[47] U. S. EPA, "Inventory of U.S. Greenhouse Gas Emissions and Sinks: 1990-2019." Available online: `https://www.epa.gov/ghgemissions/inventory-us-greenhouse-gas-emissions-and-sinks-1990-2019`, 2021. (Accessed on 30 June 2021).

[48] G. Pétron, A. Karion, C. Sweeney, B. R. Miller, S. A. Montzka, G. J. Frost, M. Trainer, P. Tans, A. Andrews, J. Kofler, D. Helmig, D. Guenther, E. Dlugokencky, P. Lang, T. Newberger, S. Wolter, B. Hall, P. Novelli, A. Brewer, S. Conley, M. Hardesty, R. Banta, A. White, D. Noone, D. Wolfe, and R. Schnell, "A new look at methane and nonmethane hydrocarbon emissions from oil and natural gas operations in the Colorado Denver-Julesburg Basin," *Journal of Geophysical Research: Atmospheres*, vol. 119, no. 11, pp. 6836–6852, 2014.

[49] A. R. Brandt, G. A. Heath, E. A. Kort, F. O'Sullivan, G. Pétron, S. M. Jordaan, P. Tans, J. Wilcox, A. M. Gopstein, D. Arent, S. Wofsy, N. J. Brown, R. Bradley, G. D. Stucky, D. Eardley, and R. Harriss, "Methane Leaks from North American Natural Gas Systems," *Science*, vol. 343, pp. 733–735, Feb. 2014.

[50] C. Frankenberg, A. K. Thorpe, D. R. Thompson, G. Hulley, E. A. Kort, N. Vance, J. Borchardt, T. Krings, K. Gerilowski, C. Sweeney, S. Conley, B. D. Bue, A. D. Aubrey, S. Hook, and R. O. Green, "Airborne methane remote measurements reveal heavy-tail flux distribution in Four Corners region," *Proceedings of the National Academy of Sciences*, vol. 113, pp. 9734–9739, Aug. 2016.

[51] D. Zavala-Araiza, D. R. Lyon, R. A. Alvarez, K. J. Davis, R. Harriss, S. C. Herndon, A. Karion, E. A. Kort, B. K. Lamb, X. Lan, A. J. Marchese, S. W. Pacala, A. L. Robinson, P. B. Shepson, C. Sweeney, R. Talbot, A. Townsend-Small, T. I. Yacovitch, D. J. Zimmerle, and S. P. Hamburg, "Reconciling divergent estimates of oil and gas methane emissions," *Proceedings of the National Academy of Sciences*, vol. 112, pp. 15597–15602, Dec. 2015.

[52] D. Zavala-Araiza, D. Lyon, R. A. Alvarez, V. Palacios, R. Harriss, X. Lan,

R. Talbot, and S. P. Hamburg, "Toward a Functional Definition of Methane Super-Emitters: Application to Natural Gas Production Sites," *Environmental Science & Technology*, vol. 49, pp. 8167–8174, July 2015.

[53] C. W. Rella, T. R. Tsai, C. G. Botkin, E. R. Crosson, and D. Steele, "Measuring Emissions from Oil and Natural Gas Well Pads Using the Mobile Flux Plane Technique," *Environmental Science & Technology*, vol. 49, pp. 4742–4748, Apr. 2015.

[54] M. Omara, M. R. Sullivan, X. Li, R. Subramanian, A. L. Robinson, and A. A. Presto, "Methane Emissions from Conventional and Unconventional Natural Gas Production Sites in the Marcellus Shale Basin," *Environmental Science & Technology*, vol. 50, pp. 2099–2107, Feb. 2016.

[55] A. M. Robertson, R. Edie, D. Snare, J. Soltis, R. A. Field, M. D. Burkhart, C. S. Bell, D. Zimmerle, and S. M. Murphy, "Variation in Methane Emission Rates from Well Pads in Four Oil and Gas Basins with Contrasting Production Volumes and Compositions," *Environmental Science & Technology*, vol. 51, pp. 8832–8840, Aug. 2017.

[56] A. R. Brandt, G. A. Heath, and D. Cooley, "Methane Leaks from Natural Gas Systems Follow Extreme Distributions," *Environmental Science & Technology*, vol. 50, pp. 12512–12520, Nov. 2016.

[57] T. L. Vaughn, C. S. Bell, C. K. Pickering, S. Schwietzke, G. A. Heath, G. Pétron, D. J. Zimmerle, R. C. Schnell, and D. Nummedal, "Temporal variability largely explains top-down/bottom-up difference in methane emission estimates from a natural gas production region," *Proceedings of the National Academy of Sciences*, vol. 115, pp. 11712–11717, Nov. 2018.

[58] S. Coburn, C. B. Alden, R. Wright, K. Cossel, E. Baumann, G.-W. Truong, F. Giorgetta, C. Sweeney, N. R. Newbury, K. Prasad, I. Coddington, and G. B. Rieker, "Regional trace-gas source attribution using a field-deployed dual frequency comb spectrometer," *Optica*, vol. 5, pp. 320–327, Apr. 2018.

[59] U. S. EPA, "Petroleum Refinery Sector Rule (Risk and Technology Review and New Source Performance Standards)." Available online: `https://www.epa.gov/stationary-sources-air-pollution/petroleum-refinery-sector-rule-risk-and-technology-review-and-new`, 2020. (Accessed on 30 June 2021).

[60] Colorado Department of Public Health and Environment, "Regulation 7: Control of Ozone via Ozone Precursors and Control of Hydrocarbons

via Oil and Gas Emissions (Emissions of Volatile Organic Compounds and Nitrogen Oxides)." Available online: https://cdphe.colorado.gov/aqcc-regulations, 2021. (Accessed on 14 July 2021).

[61] E. Yee, "Bayesian probabilistic approach for inverse source determination from limited and noisy chemical or biological sensor concentration measurements," in *Chemical and Biological Sensing VIII*, vol. 6554, p. 65540W, International Society for Optics and Photonics, Apr. 2007.

[62] R. P. Adams and D. J. C. MacKay, "Bayesian Online Changepoint Detection," *arXiv:0710.3742 [stat]*, Oct. 2007.

[63] Z. Warhaft, "Passive Scalars in Turbulent Flows," *Annual Review of Fluid Mechanics*, vol. 32, pp. 203–240, Jan. 2000.

[64] A. M. Obukhov, "Turbulence in an atmosphere with a non-uniform temperature," *Boundary-Layer Meteorology*, vol. 2, pp. 7–29, Mar. 1971.

[65] T. W. Horst and J. C. Weil, "Footprint estimation for scalar flux measurements in the atmospheric surface layer," *Boundary-Layer Meteorology*, vol. 59, pp. 279–296, May 1992.

[66] J. L. Lumley and H. A. Panofsky, *The structure of atmospheric turbulence (Interscience monographs and texts in physics and astronomy)*. New York, NY, USA: Wiley, 1964.

[67] M. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Transactions on Signal Processing*, vol. 50, pp. 174–188, Feb. 2002.

[68] H. L. Brantley, E. D. Thoma, W. C. Squier, B. B. Guven, and D. Lyon, "Assessment of Methane Emissions from Oil and Gas Production Pads using Mobile Measurements," *Environmental Science & Technology*, vol. 48, pp. 14508–14515, Dec. 2014.

[69] E. Yee, "Theory for Reconstruction of an Unknown Number of Contaminant Sources using Probabilistic Inference," *Boundary-Layer Meteorology*, vol. 127, pp. 359–394, June 2008.

[70] E. T. Jaynes, "Prior Probabilities," *IEEE Transactions on Systems Science and Cybernetics*, vol. 4, pp. 227–241, Sept. 1968.

[71] E. T. Jaynes, *Probability Theory: the Logic of Science*. Cambridge university press, 2003.

[72] A. Keats, E. Yee, and F.-S. Lien, "Bayesian inference for source determination with applications to a complex urban environment," *Atmospheric Environment*, vol. 41, pp. 465–479, Jan. 2007.

[73] E. Yee and T. K. Flesch, "Inference of emission rates from multiple sources using Bayesian probability theory," *Journal of Environmental Monitoring*, vol. 12, pp. 622–634, Mar. 2010.

[74] J. D. Wilson and B. L. Sawford, "Review of Lagrangian stochastic models for trajectories in the turbulent atmosphere," *Boundary-Layer Meteorology*, vol. 78, pp. 191–210, Feb. 1996.

[75] D. J. Thomson, "Criteria for the selection of stochastic models of particle trajectories in turbulent flows," *Journal of Fluid Mechanics*, vol. 180, pp. 529–556, July 1987.

[76] K. S. Rao, "Uncertainty Analysis in Atmospheric Dispersion Modeling," *pure and applied geophysics*, vol. 162, pp. 1893–1917, Oct. 2005.

[77] C. Forbes, M. Evans, N. Hastings, and B. Peacock, *Statistical Distributions*. Wiley, 4th ed., 2010.

[78] M. J. Flynn, "Some Computer Organizations and Their Effectiveness," *IEEE Transactions on Computers*, vol. C-21, pp. 948–960, Sept. 1972.

[79] T. A. Foster-Wittig, E. D. Thoma, and J. D. Albertson, "Estimation of point source fugitive emission rates from a single sensor time series: A conditionally-sampled Gaussian plume reconstruction," *Atmospheric Environment*, vol. 115, pp. 101–109, Aug. 2015.

[80] B. Efron and R. Tibshirani, "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy," *Statistical Science*, vol. 1, pp. 54–75, Feb. 1986.

[81] P. Das and R. Horton, "Pollution, health, and the planet: time for decisive action," *The Lancet*, vol. 391, pp. 407–408, Feb. 2018.

[82] F. Dominici, R. D. Peng, M. L. Bell, L. Pham, A. McDermott, S. L. Zeger, and J. M. Samet, "Fine Particulate Air Pollution and Hospital Admission

for Cardiovascular and Respiratory Diseases," *JAMA*, vol. 295, pp. 1127–1134, Mar. 2006.

[83] X. Wu, R. C. Nethery, B. M. Sabath, D. Braun, and F. Dominici, "Exposure to air pollution and COVID-19 mortality in the United States: A nationwide cross-sectional study," *medRxiv*, p. 2020.04.05.20054502, Apr. 2020.

[84] W. Q. Gan, H. W. Davies, M. Koehoorn, and M. Brauer, "Association of Long-term Exposure to Community Noise and Traffic-related Air Pollution With Coronary Heart Disease Mortality," *American Journal of Epidemiology*, vol. 175, pp. 898–906, May 2012.

[85] S. Shin, L. Bai, T. H. Oiamo, R. T. Burnett, S. Weichenthal, M. Jerrett, J. C. Kwong, M. S. Goldberg, R. Copes, A. Kopp, and H. Chen, "Association Between Road Traffic Noise and Incidence of Diabetes Mellitus and Hypertension in Toronto, Canada: A Population-Based Cohort Study," *Journal of the American Heart Association*, vol. 9, p. e013021, Mar. 2020.

[86] Y. Ogen, "Assessing nitrogen dioxide (NO2) levels as a contributing factor to coronavirus (COVID-19) fatality," *Science of The Total Environment*, vol. 726, p. 138605, July 2020.

[87] G. S. W. Hagler, M.-Y. Lin, A. Khlystov, R. W. Baldauf, V. Isakov, J. Faircloth, and L. E. Jackson, "Field investigation of roadside vegetative and structural barrier impact on near-road ultrafine particle concentrations under a variety of wind conditions," *Science of The Total Environment*, vol. 419, pp. 7–15, Mar. 2012.

[88] J. S. Apte, K. P. Messier, S. Gani, M. Brauer, T. W. Kirchstetter, M. M. Lunden, J. D. Marshall, C. J. Portier, R. C. Vermeulen, and S. P. Hamburg, "High-Resolution Air Pollution Mapping with Google Street View Cars: Exploiting Big Data," *Environmental Science & Technology*, vol. 51, pp. 6999–7008, June 2017.

[89] P. Deshmukh, S. Kimbrough, S. Krabbe, R. Logan, V. Isakov, and R. Baldauf, "Identifying air pollution source impacts in urban communities using mobile monitoring," *Science of The Total Environment*, vol. 715, p. 136979, May 2020.

[90] C. A. Pope, "Epidemiology of fine particulate air pollution and human health: biologic mechanisms and who's at risk?," *Environmental Health Perspectives*, vol. 108, pp. 713–723, Aug. 2000.

[91] R. Morello-Frosch, M. Pastor, and J. Sadd, "Environmental Justice and Southern California's "Riskscape": The Distribution of Air Toxics Exposures and Health Risks among Diverse Communities," *Urban Affairs Review*, vol. 36, pp. 551–578, Mar. 2001.

[92] G. A. Millett, A. T. Jones, D. Benkeser, S. Baral, L. Mercer, C. Beyrer, B. Honermann, E. Lankiewicz, L. Mena, J. S. Crowley, J. Sherwood, and P. Sullivan, "Assessing Differential Impacts of COVID-19 on Black Communities," *Annals of Epidemiology*, pp. 37–44, May 2020.

[93] G. S. W. Hagler, E. D. Thoma, and R. W. Baldauf, "High-Resolution Mobile Monitoring of Carbon Monoxide and Ultrafine Particle Concentrations in a Near-Road Environment," *Journal of the Air & Waste Management Association*, vol. 60, pp. 328–336, Mar. 2010.

[94] D. Hasenfratz, O. Saukh, C. Walser, C. Hueglin, M. Fierz, T. Arn, J. Beutel, and L. Thiele, "Deriving high-resolution urban air pollution maps using mobile sensor nodes," *Pervasive and Mobile Computing*, vol. 16, pp. 268–285, Jan. 2015.

[95] J. Wallace, D. Corr, P. Deluca, P. Kanaroglou, and B. McCarry, "Mobile monitoring of air pollution in cities: the case of Hamilton, Ontario, Canada," *Journal of Environmental Monitoring*, vol. 11, pp. 998–1003, May 2009.

[96] C. E. Kolb, S. C. Herndon, J. B. McManus, J. H. Shorter, M. S. Zahniser, D. D. Nelson, J. T. Jayne, M. R. Canagaratna, and D. R. Worsnop, "Mobile Laboratory with Rapid Response Instruments for Real-Time Measurements of Urban and Regional Trace Gas and Particulate Distributions and Emission Source Characteristics," *Environmental Science & Technology*, vol. 38, pp. 5694–5703, Nov. 2004.

[97] H. L. Brantley, G. S. W. Hagler, E. S. Kimbrough, R. W. Williams, S. Mukerjee, and L. M. Neas, "Mobile air monitoring data-processing strategies and effects on spatial air pollution trends," *Atmospheric Measurement Techniques*, vol. 7, pp. 2169–2183, July 2014.

[98] M. Van Poppel, J. Peters, and N. Bleux, "Methodology for setup and data processing of mobile air quality measurements to assess the spatial variability of concentrations in urban environments," *Environmental Pollution*, vol. 183, pp. 224–233, Dec. 2013.

[99] L. M. Zwack, C. J. Paciorek, J. D. Spengler, and J. I. Levy, "Modeling Spa-

tial Patterns of Traffic-Related Air Pollutants in Complex Urban Terrain," *Environmental Health Perspectives*, vol. 119, pp. 852–859, June 2011.

[100] R. Hagemann, U. Corsmeier, C. Kottmeier, R. Rinke, A. Wieser, and B. Vogel, "Spatial variability of particle number concentrations and NOx in the Karlsruhe (Germany) area obtained with the mobile laboratory 'AERO-TRAM'," *Atmospheric Environment*, vol. 94, pp. 341–352, Sept. 2014.

[101] Environmental Defense Fund, "Why new technology is critical for tackling air pollution around the globe." `https://www.edf.org/airqualitymaps`. Accessed February 20, 2021.

[102] J. L. Pearce, J. Beringer, N. Nicholls, R. J. Hyndman, and N. J. Tapper, "Quantifying the influence of local meteorology on air quality using generalized additive models," *Atmospheric Environment*, vol. 45, pp. 1328–1336, Feb. 2011.

[103] J. P. Dawson, P. J. Adams, and S. N. Pandis, "Sensitivity of ozone to summertime climate in the eastern USA: A modeling case study," *Atmospheric Environment*, vol. 41, pp. 1494–1511, Mar. 2007.

[104] K. P. Messier, S. E. Chambliss, S. Gani, R. Alvarez, M. Brauer, J. J. Choi, S. P. Hamburg, J. Kerckhoffs, B. LaFranchi, M. M. Lunden, J. D. Marshall, C. J. Portier, A. Roy, A. A. Szpiro, R. C. H. Vermeulen, and J. S. Apte, "Mapping Air Pollution with Google Street View Cars: Efficient Approaches with Mobile Monitoring and Land Use Regression," *Environmental Science & Technology*, vol. 52, pp. 12563–12572, Nov. 2018.

[105] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-Means Clustering Algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[106] Google, "Raw air quality data from google / aclima." `https://goo.gl/q4TRtt`. Accessed November 1, 2020.

[107] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. v. d. Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, I. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, and P. v. Mulbregt, "SciPy 1.0: fundamental algorithms for scientific computing in Python," *Nature Methods*, vol. 17, pp. 261–272, Mar. 2020.

[108] D. Roberts–Semple, F. Song, and Y. Gao, "Seasonal characteristics of ambient nitrogen oxides and ground–level ozone in metropolitan northeastern New Jersey," *Atmospheric Pollution Research*, vol. 3, pp. 247–257, Apr. 2012.

[109] R. E. Britter and S. R. Hanna, "Flow and Dispersion in Urban Areas," *Annual Review of Fluid Mechanics*, vol. 35, no. 1, pp. 469–496, 2003.

[110] E. S. Kimbrough, R. W. Baldauf, and N. Watkins, "Seasonal and diurnal analysis of NO2 concentrations from a long-duration study conducted in Las Vegas, Nevada," *Journal of the Air & Waste Management Association (1995)*, vol. 63, pp. 934–942, Aug. 2013.

[111] J. Richmond-Bryant, M. Snyder, R. Owen, and S. Kimbrough, "Factors associated with NO2 and NOX concentration gradients near a highway," *Atmospheric environment*, vol. 174, pp. 214–226, Nov. 2017.

[112] R. W. Atkinson, B. K. Butland, H. R. Anderson, and R. L. Maynard, "Long-term Concentrations of Nitrogen Dioxide and Mortality," *Epidemiology*, vol. 29, pp. 460–472, July 2018.

[113] D. J. Briggs, S. Collins, P. Elliot, P. Fischer, S. Kingham, E. Lebret, K. Pryl, H. V. Reeuwijk, K. Smallbone, and A. V. D. Veen, "Mapping urban air pollution using GIS: a regression-based approach," *International Journal of Geographical Information Science*, vol. 11, pp. 699–718, Oct. 1997.

[114] M. Jerrett, A. Arain, P. Kanaroglou, B. Beckerman, D. Potoglou, T. Sahsuvaroglu, J. Morrison, and C. Giovis, "A review and evaluation of intraurban air pollution exposure models," *Journal of Exposure Science & Environmental Epidemiology*, vol. 15, pp. 185–204, Mar. 2005.

[115] X. Xie, I. Semanjski, S. Gautama, E. Tsiligianni, N. Deligiannis, R. T. Rajan, F. Pasveer, and W. Philips, "A Review of Urban Air Pollution Monitoring and Exposure Assessment Methods," *ISPRS International Journal of Geo-Information*, vol. 6, p. 389, Dec. 2017.

[116] C. Ding, X. He, H. Zha, and H. Simon, "Adaptive dimension reduction for clustering high dimensional data," Tech. Rep. LBNL-51472, Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States), Oct. 2002.

[117] D. Arthur and S. Vassilvitskii, "k-means++: the advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, SODA '07, (USA), pp. 1027–1035, Society for Industrial and Applied Mathematics, Jan. 2007.

[118] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer Series in Statistics, New York, NY: Springer New York, 2009.

[119] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, no. 2, pp. 411–423, 2001.

[120] R. J. Delfino, R. S. Zeiger, J. M. Seltzer, D. H. Street, and C. E. McLaren, "Association of asthma symptoms with peak particulate air pollution and effect modification by anti-inflammatory medication use.," *Environmental Health Perspectives*, vol. 110, pp. A607–A617, Oct. 2002.

[121] R. Vautard, J. Cattiaux, P. Yiou, J.-N. Thépaut, and P. Ciais, "Northern Hemisphere atmospheric stilling partly attributed to an increase in surface roughness," *Nature Geoscience*, vol. 3, pp. 756–761, Nov. 2010.

[122] D. Belušić, D. H. Lenschow, and N. J. Tapper, "Performance of a mobile car platform for mean wind and turbulence measurements," *Atmospheric Measurement Techniques*, vol. 7, pp. 1825–1837, June 2014.

[123] A. N. Kolmogorov, "Dissipation of Energy in Locally Isotropic Turbulence," *Akademiia Nauk SSSR Doklady*, vol. 32, p. 16, Apr. 1941.

[124] C.-I. Hsieh and G. G. Katul, "Dissipation methods, Taylor's hypothesis, and stability correction functions in the atmospheric surface layer," *Journal of Geophysical Research: Atmospheres*, vol. 102, no. D14, pp. 16391–16405, 1997. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/97JD00200.

[125] T. Duman, G. G. Katul, M. B. Siqueira, and M. Cassiani, "A Velocity–Dissipation Lagrangian Stochastic Model for Turbulent Dispersion in Atmospheric Boundary-Layer and Canopy Flows," *Boundary-Layer Meteorology*, vol. 152, pp. 1–18, July 2014.

[126] H. Rodean, *Stochastic Lagrangian Models of Turbulent Diffusion*. American Meteorological Society, 1st ed., 1996.

[127] C.-I. Hsieh, G. Katul, and T.-w. Chi, "An approximate analytical model for footprint estimation of scalar fluxes in thermally stratified atmospheric flows," *Advances in Water Resources*, vol. 23, pp. 765–772, June 2000.

[128] T. K. Flesch, J. D. Wilson, and E. Yee, "Backward-Time Lagrangian

Stochastic Dispersion Models and Their Application to Estimate Gaseous Emissions," *Journal of Applied Meteorology and Climatology*, vol. 34, pp. 1320–1332, June 1995. Publisher: American Meteorological Society Section: Journal of Applied Meteorology and Climatology.

[129] U. Rannik, A. Sogachev, T. Foken, M. Gockede, N. Kljun, M. Y. Leclerc, and T. Vesala, *Footprint analysis. In: Eddy Covariance: A Practical Guide to Measurement and Data Analysis.* Springer, 2012.

[130] City of Oakalnd, "See the City Zoning Map." Available online: `https://www.oaklandca.gov/resources/zoning-map`, 2020. (Accessed on 1 November 2020).

[131] U.S. EPA, "Superfund: National Priorities List (NPL)." Available online: `https://www.epa.gov/superfund/superfund-national-priorities-list-npl`, 2020. (Accessed on 15 October 2020).

[132] U.S. EPA, "Find, Understand and Use TRI." Available online: `https://www.epa.gov/toxics-release-inventory-tri-program/find-understand-and-use-tri`, 2020. (Accessed on 15 October 2020).

[133] T. G. Farr, P. A. Rosen, E. Caro, R. Crippen, R. Duren, S. Hensley, M. Kobrick, M. Paller, E. Rodriguez, L. Roth, D. Seal, S. Shaffer, J. Shimada, J. Umland, M. Werner, M. Oskin, D. Burbank, and D. Alsdorf, "The Shuttle Radar Topography Mission," *Reviews of Geophysics*, vol. 45, no. 2, 2007. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2005RG000183.

[134] C. E. Woodcock, R. Allen, M. Anderson, A. Belward, R. Bindschadler, W. Cohen, F. Gao, S. N. Goward, D. Helder, E. Helmer, R. Nemani, L. Oreopoulos, J. Schott, P. S. Thenkabail, E. F. Vermote, J. Vogelmann, M. A. Wulder, and R. Wynne, "Free Access to Landsat Imagery," *Science*, vol. 320, pp. 1011–1011, May 2008.

[135] A. Polhamus, J. B. Fisher, and K. P. Tu, "What controls the error structure in evapotranspiration models?," *Agricultural and Forest Meteorology*, vol. 169, pp. 12–24, Feb. 2013.