Integrating Data Analytics and Mass Balance Approaches to Estimate and Understand Regional Methane Emissions: A Study in the Permian Basin

by

Yifan Bian

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Petroleum Engineering

Department of Civil & Environmental Engineering
University of Alberta

ABSTRACT

Satellite-retrieved methane (CH_4) concentration data offers a valuable opportunity for large-scale emissions monitoring. However, its widespread adoption remains challenging due to the data volume and varying data quality. A workflow to estimate the methane emission rate of major hydrocarbon plays based on the mass balance principle using publicly available Sentinel-5P satellite data is presented. This workflow estimates the methane emission rate originating from specific regions. The proposed workflow is applied to estimate emissions from the Permian and Appalachian Basins in the United States. Applying the proposed workflow to these regions, the three-year-mean methane emission rates from 2019 to 2021 are estimated to be 3.56 Mt/year for the Permian Basin and 4.46 Mt/year for the Appalachian Basin. The results are compared against volumes estimated by other means and reported in the literature. The proposed method is easy to implement and offers promising potential for practical and reliable estimates for long-term regional methane emission monitoring purposes for operators, governments, investors, and the general public. In addition, this study presents a comprehensive, data-driven approach to analyze and predict methane enhancements in the Permian Basin. Methane enhancement refers to "*the increase in methane concentration above the baseline background level*" (Dlugokencky et al., 2003). Leveraging satellite-retrieved methane (CH_4) concentration data and oil and gas related operational data, this research helps to better understand the complex interactions influencing methane emissions. It begins with a descriptive analysis of methane enhancement data attributing to different operators based on their geographical distribution across the basin. Next, multiple supervised and unsupervised learning algorithms are utilized to help predict methane enhancement levels quantitatively, offering insights into influential features contributing to methane emissions. Lastly, impurity-based feature importance and SHAP values are used to evaluate the predictive power and interpretability of these models, decoding the 'black-box' nature and enabling an in-depth understanding of the factors driving methane enhancements. This study explores the complex dynamics of methane emissions in the Permian Basin but also sets a foundation for future investigations aimed at refining our comprehension and prediction capabilities of methane emissions in oil and gas regions.

PREFACE

This thesis is original work by Yifan Bian. Some parts in chapter 1, 2 and 3 of this thesis have been published or presented as:

Bian, Jeffrey Y., Leung, Juliana Y., Volkmer, Nick, and Jingwen Zheng. "An Improved Workflow in Mass Balance Approach for Estimating Regional Methane Emission Rate Using Satellite Measurements." Paper presented at the SPE Canadian Energy Technology Conference and Exhibition, Calgary, Alberta, Canada, March 2023. Doi: https://doi.org/10.2118/212791-MS

DEDICATION

Dedicated to my family, my friends, and my teachers for their love, endless support and mentorship.

ACKNOWLEDGMENTS

Firstly, I would like to extend my deepest gratitude to my supervisor, Dr Juliana Y. Leung for her continuous support and guidance throughout my studies and research. This thesis would not have been possible without her guidance.

I am grateful to Nick Volkmer and Jingwen Zheng's team at Enverus for their insightful comments and suggestions throught the research.

My appreciation goes to Mitacs for their financial support through the Mitacs Accelerate program (IT28455).

I would also like to thank the collaboration between Enverus and the University of Alberta and appreciate the data usage from Enverus PRISM.

Finally, I would like to give my eternal gratitude to my parents and my partner Weijia. I could not have gotten this far without their endless love, support and encouragement.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 Background

Methane, also known as CH4, is a potent greenhouse gas (GHG) with a significant environmental impact. It has a global warming potential of more than 80 times greater than carbon dioxide (CO_2) over a 20-year period. This high potency makes methane more effective in trapping heat in the atmosphere than other GHGs, making it a critical issue to address. In addition to being a potent GHG, methane constitutes an essential component of natural gas. To achieve the goal of net Zero Emissions by 2050 set by the United Nations (2022), reducing methane emissions from oil and gas operations is one of the most cost-effective and impactful actions governments can take. Oil and gas operations are among the largest anthropogenic sources of methane emissions, with agricultural emissions being the other dominant source. Recognizing the need to address methane emissions has led many countries and regions to take steps toward reducing emissions through regulatory and voluntary industry actions. This global mission underscores the importance of considering methane emissions as a critical factor in meeting short-term global climate goals.

The first step is to utilize a cost-effective resource to monitor and estimate methane emissions. BloombergNEF (2022) estimated that there would be a projected market size of around 900 million USD by 2025 in methane monitoring within the oil and gas industry. Approximately 85% of the budget will be allocated toward upstream activities, where most emissions originate. For example, methane emissions can originate from various upstream and downstream sources, such as fluid flowbacks during completion, unloading liquids, pneumatic devices, pumps, workovers involving hydraulic fracturing, separator systems, pumps, storage tanks, and various onsite equipment, pipeline connections and processing plants. Most of these emissions occur through atmospheric venting and controlled flaring due to inadequate proper gathering and boosting systems. Some fugitive leaking events could also be unexpected ultra-emission sources.

Existing methane monitoring approaches vary from stationary sensors for point source monitoring (Kumar et al., 2022), drones for areal monitoring (Tuzson et al., 2020), and aircraft surveys for regional monitoring (Yakovlev et al., 2022), to satellites for large-scale or global monitoring. While field measurement of methane (bottom-up method) can provide an in-depth understanding and identify local point sources within a small area, it would require tremendous effort and resources to apply continuously and consistently over a larger area. On the other hand, atmospheric measurements of methane concentration using satellites (top-down method) are more applicable and more efficient in terms of cost in monitoring regional methane emission status/trends and enable timely identification of fugitive methane emission compared to bottom-up methods and top-down methods using drones or aircrafts. Another advantage of leveraging satellite data is its accessibility. Many methane-tracking satellites are either free to the public or accessible at a very low cost compared to the other three means of tracking methane emissions.

In recent years, the advent of satellite remote sensing technology, such as the Sentinel-5 Precursor (S5p) satellite and its Tropospheric Monitoring Instrument (TROPOMI), can measure methane concentrations on a global scale with high revisit time of less than one day (Sentinel hub, 2023). Here we demonstrate the capability of quantifying methane emission using the measurements from S5p. The S5p mission is designed to measure the Near Infrared (NIR) and Short-Wave Infrared (SWIR) spectral range daily (Hu et al., 2016), providing a unique opportunity to gather consistent and accurate methane measurement data. The CH4 absorption lines in the SWIR band are used to retrieve methane columns using the RemoTec algorithm (Butz et al., 2009), which is based on a full-physics approach and was developed using data from the Greenhouse Gases Observing SATellite (GOSAT).

1.2 Problem Statement

The ability to monitor methane emissions on a large scale is a critical requirement for operators, governments, investors, and the general public to manage environmental impact and adhere to sustainable practices. The data for such monitoring can be sourced from puiblicly available satellites such as the Sentinel-5P. However, several challenges obstruct its widespread adoption, including the immense colume of satellite data, inconsistent data quality, and discrepancies existing between different measurement methodologies (e.g. top-down measurements sourced froms arellites vs bottom-up measurements). A systematic workflow is needed to estimate methane emission rates from large hydrocarbon fields, such as the Pemrian and the Appalachian Basins in the United States.

Additionally, the understanding of methane enhancements- the increase in methane concentration above the baseline background level- and the complex interactions leading to these enhancements in regions like the Permian basin – the largest oil-producing basin in the US, is currently limited. The existing literatures lacks a comprehensive, data driven approach to analyze and correlate these methane enhancements, which could be crucial for understanding the complexity of methane emission sources in the oil and gas industry.

This research aims to address these gaps and limitations in the current understanding and methodologies of methane emission monitoring and correlation. It also aims to lay the groundwork for future studies focused on improving the estimation, comprehension, and predictive capabilities for methane emissions in oil and gas producing regions using satellite measurements.

1.3 Research Objective

This theme of this research can be divided into the following objectives:

1. Develop a workflow to quantify regional methane emission rates that is straightforward, easy to implement, and holding promise for providing practical long-term methane emission monitoring using the Sentinel-5P data. It can be achieved by:
    1) Propose a series of operations to preprocess the retrieved Sentinel-5P data.
    2) Develop a physics-based method to quantify the rate of methane emission using the processed Sentinel-5P data.
2. Demonstrate the feasibility of proposed workflow in real-world application, including estimating methane emission rates in the Permian Basin and the Appalachian Basin.
3. Develop a comprehensive understanding of the factors that contribute to methane emissions in the oil and gas industry. It can be achieved by:
    1) Utilizing machine learning techniques to analyze a large dataset comprising Sentinel-5P TROPOMI methane concentration data and oil and gas operating parameters such as the number of wells, production volumes, well types, and operator groups, among others.
    2) Identifying key drivers of methane emissions by establishing correlations or between variables and evaluating their influences and contributions in the model.

1.4 Structure of Thesis

This thesis consists of 8 chapters, and it is organized as follows:

In Chapter 1, we introduce the topic and outline the background of the research problem.

Chapter 2 reviews previous literatures including a history of satellite measurements on methane emissions, studies on methane emission with Sentinel-5P satellite, proposed methodologies which can be used to estimate rate of emission of methane.

In Chapter 3, we focus on the estimation of methane emission rates, describing the methods employed for the retrieval and preprocessing of Sentinel-5P TROPOMI products in this study and literature review on existing studies related to methane emission estimation methodologies, as well as the algorithms for monthly and daily methane emission calculations and rate of emission estimation. The results of our analysis are presented in Section 3.3, focusing on two major regions: the Permian Basin and the Appalachian Basin, followed by a discussion of the uncertainties involved in our estimations.

Chapter 4 delves into the correlation of methane enhancement with oil and gas parameters, including operator emission allocation, predictive models, and emission flagging techniques.

Finally, in Chapter 5, conclusions from this research are summarized, highlighting the key findings and their implications for the oil and gas industry. Overall, this thesis aims to provide a comprehensive understanding of the factors influencing methane emissions in the oil and gas industry and offer insights that could inform targeted mitigation strategies.

CHAPTER 2: LITERATURE REVIEW

2.1 Satellite Measurements on Methane Emissions

The first attempt at observing atmospheric methane from space began in the late 20th century. In the late 1990s, the European Space Agency's (ESA) Global Ozone Monitoring Experiment (GOME) satellite was launched to monitor atmospheric ozone, but it also provided data on methane. These measurements, while groundbreaking, were limited in spatial resolution (about 40 × 2 km^2 on the earth’s surface)(Breiman at al., 1999). Entering the 21st century, the SCanning Imaging Absorption spectroMeter for Atmospheric CHartographY (SCIAMACHY) was a mission, launched by ESA in 2002, that carried an instrument on the Envisat satellite. SCIAMACHY provided measurements of methane with a resolution far superior to the previous generation of satellites. However, it had limitations in coverage and sensitivity (Frankenberg et al., 2005). 7 years later, the Japanese Greenhouse gases Observing SATellite (GOSAT), launched in 2009, was the first satellite specifically designed to measure greenhouse gases, including methane. It used a different detection method, namely, Fourier Transform Spectrometry, which enabled more precise measurements than previous satellites (Kuzu et al., 2016). In 2017, ESA launched the Tropospheric Monitoring Instrument (TROPOMI) aboard the Sentinel-5P satellite. TROPOMI provided daily global methane measurements with unprecedented spatial resolution of 7.0 x 5.5 km^2 (Hasekamp et al., 2019).

In recent years, there have been significant advancement in the private sector. Companies like GHGSat have developed and launched their satellites specifically to detect and quantify greenhouse gas emissions including methane with a spatial resolution of up to 25m, from individual industrial sites (GHGSat., 2021). In addition, plans are underway for the launch of further advanced methane monitoring missions, such as ESA's Copernicus Anthropogenic Carbon Dioxide Monitoring (CO2M) mission (Sierk et al., 2021) and others.

2.2 Applications using Sentinel-5P Measurements

Multiple studies have been conducted to evaluate the ability of S5p products to estimate methane emissions on a regional scale (Varon et al., 2018, 2019; Lyon et al., 2021; de Gouw et al., 2020; Pankaj et al., 2020; Zhang et al., 2020; Sadavarte et al., 2021; Schneising et al., 2020, Shen et al., 2022; Pandey et al., 2019), and these studies have consistently shown the high potential of the S5p and TROPOMI for this application.

For example, Zhang et al. (2020) applied atmospheric inverse modeling on TROPOMI observations to estimate monthly methane emissions for the Permian Basin. This model uses the relationship between the observed methane enhancements and the underlying emissions, taking into account atmospheric transport and chemistry. The model was run using the GEOS-Chem chemical transport model, which simulates the transport and chemistry of atmospheric gases around the globe. The authors used a high-resolution version of the model (0.25° × 0.3125°) to accurately simulate the transport of methane in the Permian Basin. The authors found that methane emissions from oil and natural gas production in the Permian Basin are estimated to be 2.7 ± 0.5 Tg per year from May 2018 to March 2019, They also compared the emissions estimated from the inverse model with those reported in the EPA's Greenhouse Gas Inventory. They found that the inverse model estimates were more than twice as high as the inventory estimates or ~60% higher thane the national average methane leakage rate, suggesting that the inventory may be underestimating emissions from the Permian Basin. The authors also suggested that the high leakage rate in the Permian Basin seems to be linked with a lack of adequate infrastructure for the collection, processing, and transportation of natural gas, resulting in widespread venting and flaring.

In the application of point quantification, Pandey et al. (2019) used the mass balance approach utilizing WRF simulation to estimate methane emissions from a natural gas well blowout in the Appalachian Basin. This is a mathematical technique that simulates methane measurements at various locations and times to estimate the emissions that would have led to those TROPOMI measurements. The inversion model used in this study was based on the Weather Research and Forecasting model coupled with Chemistry (WRF-Chem), a widely used model for simulating the transport and transformation of gases in the atmosphere. These modelling approaches were impossible with previous generations of methane-measuring satellites, like GOSAT or SCIAMACHY, due to their limited temporal and spatial resolution. The authers also used cross-sectional flux method to quantify the emission (Varon, 2018). Both methods yielded similar results, indicating a methane emission rate of approximately 120 ± 32 metric tons per hour. This emission rate was then used to estimate the total methane emission during the 20-day blowout period, resulting in a total of 60 ± 15 kilotons.

Schneising et al. (2020) utilized daily measurements of TROPOMI to estimate regional methane emission rate using a method based on mass balance under a rotated coordinate system so that the wind directions were homogenized in the region. After rotating the coordinate system, the transformed daily data was gridded on a 0.05° × 0.05° grid. Grid boxes with methane concentrations below the 10th percentile within a radius of 700 km around the pivotal point were excluded due to potential residual cloud cover. After subtracting a suitable background, the data was used to estimate the daily emission rate. This was done by calculating fluxes of the vector field through cross-sections perpendicular to wind direction according to the divergence theorem. The authors found that the mean emission estimate for the period 2018/2019 was 3.16±1.13 Mt per year. This corresponds to a fugitive emission rate of 1.3±0.5% relative to combined oil and gas energy production.

Chapter 3 utilized the approach proposed by Schneising et al. (2020), but without coordinate rotation, in order to preserve the native geographical coordinate system while accommodating the varying wind directions. A method of interpolation has been integrated into the workflow to maximize the utilization of high-quality S5p observations and avoid discarding a significant number of them. By incorporating this approach, the methane emission rate can be determined with greater frequency, allowing for more timely monitoring of methane emissions daily, weekly, or monthly, as opposed to only annually. By leveraging this high revisit frequency and the precise methane measurement data gathered by S5p, it is possible to build a comprehensive and accurate picture of estimated methane emission rates worldwide. These estimates can inform and guide mitigation efforts aimed at reducing greenhouse gas emissions and achieving net zero emissions by 2050.

CHAPTER 3: METHANE EMISSION RATE ESTIMATION

3.1 Methods

3.1.1 Retrieval of Preprocessing of S5p Products

S5p products are delivered without a fixed grid (level-2 data) – the data pixels are described by geo-gridded latitude and longitude that form an irregular grid because of the polar orbit. Furthermore, every satellite orbit shifts slightly towards the east daily, so the level-2 ground pixel coverage is marginally different. The satellite passes over an identical geographic region every 16 days (Hasekamp et al., 2019). In other words, the orbit resets itself every 16 days. An irregular grid system makes combining and comparing grid measurements on different days challenging and problematic. Therefore, in this study, the S5p products are transformed from level-2 data to level-3 data by resampling the data to a regular spatial grid system using an algorithm based on the nearest-neighbor interpolation principle. One can also use the HARP toolbox (HARP 1.17) for this purpose.

The spatial resolution of the S5p methane product has been 7.0 km × 7.0 km since the mission's launch in October 2017. It was then improved to a higher resolution of 7.0 km × 5.5 km on August 6, 2019 (Hasekamp et al., 2019). To maximize data intensity from the S5p products, the transformed grid resolution is interpolated to 0.05° latitude × 0.05° longitude. As an example, in one of the main studied domains: the Permian Basin and surrounding regions in western Texas and southeastern New Mexico of the United States (30° to 34.5° N and 100° to 105° W), the uniform 0.05° × 0.05° grid size is approximately 5.5 km × 4.8 km, a slightly finer resolution than the S5p methane product.

Bias-corrected column-averaged dry-air mole fraction of methane data ($XCH4$) in the unit of parts per billion (ppb) from one of the products of S5p is used. The bias correction is performed based on the retrieved surface albedo (spectrum intensity above a threshold level) that improves the accuracy of this product (measurement). Several screening criteria are also performed to ensure the reliability of the retrieved products; only measurements collected under certain conditions are considered: (1) at the dayside of the orbit, (2) over land, (3) cloud-free, (4) solar zenith angle < 70°, (5) instrument zenith angle < 60°. Next, additional screening is performed based on the q_a

(data quality) value, one of the output products of S5p; it summarizes several quality assurance parameters (cloud fraction, terrain roughness, spectral offset, aerosol threshold etc.) into a value ranging from 0 to 1 (Apituley et al., 2017). The data products are filtered again to exclude data with a qa_value < 0.5. Only a portion of the retrieved data is selected for further application for the reasons above. Two approaches in estimating monthly or annual methane emission rates for large geographic regions are explored: (1) using daily retrieved methane columns (i.e., daily estimation) and (2) using monthly aggregated methane columns (i.e., monthly estimation).

3.1.2 Monthly Methane Interpolation

The monthly methane emission estimation is obtained by temporally aggregating daily retrieved XCH_4 data over the period of a month for every grid cell and computing its average. There are always grid cells with no measurements available, even after the monthly aggregation step, due to the filtering of low-quality data. Accordingly, to ensure sufficient data is available for estimating the monthly emission rate over a source region, an interpolation technique is used to fill in missing grid cells following temporal aggregation. One of the deterministic spatial interpolation methods, inverse distance weighted interpolation (IDW) (Shepard, 1968), is applied. Other (geo)statistical interpolation techniques, like kriging, can also be used, as in many remote sensing applications (Papritz et al., 2002). This paper uses the IDW technique, which is relatively easy to interpret and computationally efficient. The estimated value at an unsampled location is the weighted sum of all the neighboring data points. The weight assigned to each data point is d^{-p}, where d is the distance between the unsampled location and the individual data point, and p is the power parameter. Greater values of p would assign more weight to data values closest to the unsampled points. This application sets a value of 2 for p. Figure 1 illustrates the workflow of aggregating the transformed level-3 daily S5p filtered methane measurements into a monthly averaged methane enhancement heatmap. Methane enhancement refers to "*the increase in methane concentration above the baseline background level*" (Dlugokencky et al., 2003). Then, the heatmap is interpolated to fill in missing values caused by cloud cover or other factors that may have prevented measurements over the entire month. The resulting interpolated map provides a more complete and accurate representation of methane concentration in the Permian basin. Further details on methane enhancement are discussed in the next section of the thesis.

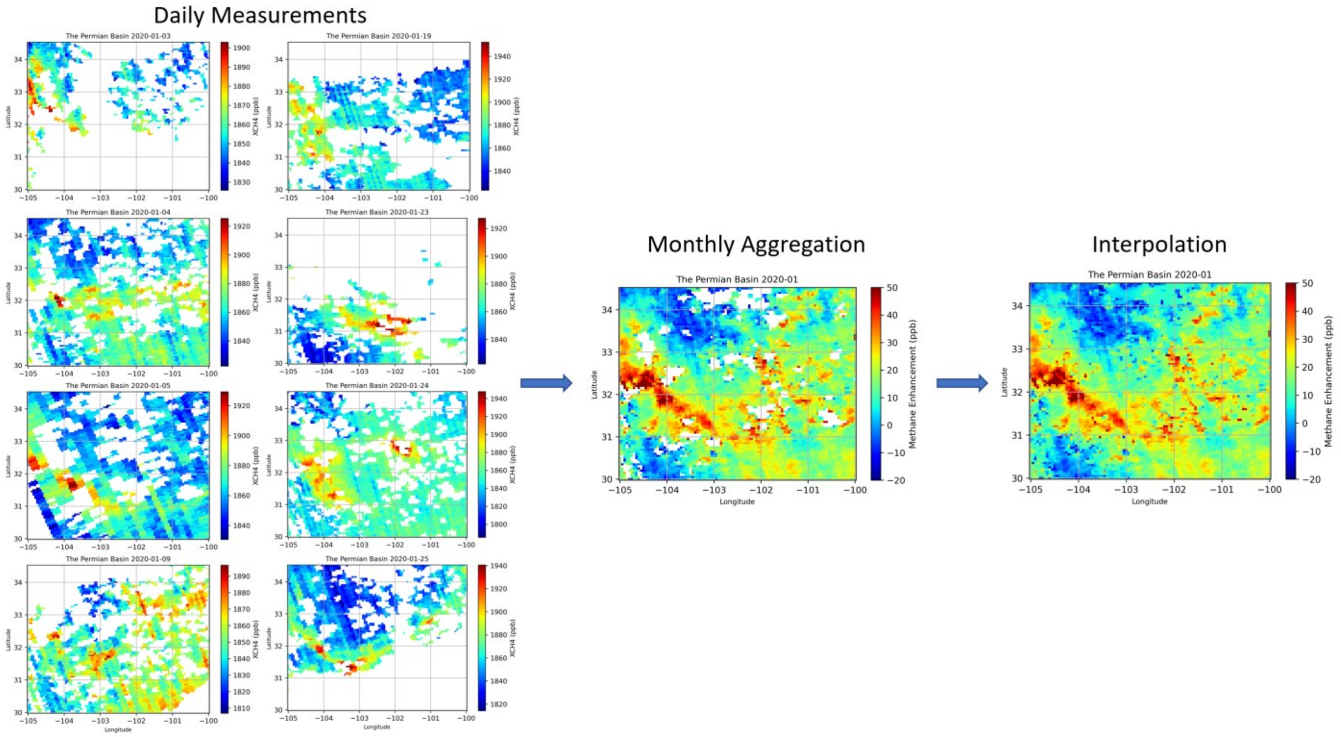

Daily Measurements
The Permian Basin 2020-01-03
The Permian Basin 2020-01-19
The Permian Basin 2020-01-04
The Permian Basin 2020-01-23
The Permian Basin 2020-01-05
The Permian Basin 2020-01-24
The Permian Basin 2020-01-09
The Permian Basin 2020-01-25
XCH4 (ppb)
Latitude
Longitude
Monthly Aggregation
The Permian Basin 2020-01
Methane Enhancement (ppb)
Interpolation
The Permian Basin 2020-01
Methane Enhancement (ppb)

Figure 1. Illustration of methane enhancement ΔXCH_4 map monthly aggregation and interpolation for the Permian basin over the month of January 2020: Left: Daily Measurements – transformed Level-3 daily S5p filtered methane measurements XCH_4. Middle: Monthly Aggregation – methane enhancement heatmap after monthly aggregation. Right: Interpolation – methane enhancement heatmap after interpolation.

3.1.3 Daily Methane Interpolation

The practice of estimating monthly emissions provides a more consistent and detailed temporal analysis of methane levels compared to the annual emission reports of major oil and gas production regions released by the U.S. Environmental Protection Agency (EPA) and other energy regulators in North America (EPA, 2023). However, the monthly aggregated XCH4 levels may be too smooth. For example, in instances where only one usable measurement is available at a grid cell, it is assumed that the XCH4 persists at the level of that single measurement for the entire month, which may result in over- or under-estimation. In principle, TROPOMI can record measurements as frequently as daily. The lack of daily data is primarily due to poor data quality and certain environmental conditions such as cloudiness (Hasekamp et al., 2019). For example, out of the 341 non-empty *XCH_4* retrievals over the Permian basin in 2019, only 20% of the retrievals (locations) have over 50% data coverage after data quality filtering, and approximately 67% of retrievals have less than 40% coverage. The data coverage is even worse in those regions with more cloudy days, e.g., the Appalachian basin.

An alternative approach is implementing a temporal interpolation scheme to estimate daily emissions. Here is the proposed workflow: First, daily retrievals are spatially interpolated based on IDW at grid cells with missing data within four grid cells from the closest true observation. Next, temporal interpolation is applied for those grid cells still missing a value after the spatial interpolation. Each unfilled grid cell is linearly interpolated using its values in the most recent past and the nearest future. It should be noted that the minimum detection limits of daily measurements from a single overpass (500-8800kg/h/pixel) are generally higher than measurements over a longer campaign (50-1200kg/h/pixel for a yearlong campaign) (Dubey et al., 2023). The quantity of emissions that can be detected also increases significantly with a longer length of measurement campaign (Dubey et al., 2023). In the case when two S5p's orbit's coverages have overlapped slightly, those grid cells within the overlapping zones would record two measurements from consecutive overpasses on the same day. In those instances, the average of those two measurements is assigned as the XCH_4 for that day.

3.1.4 Rate of Emission Estimation Algorithm

The algorithm is based on a mass balance of XCH_4 over a controlled volume V – the emission rate is equal to the cross-sectional fluxes in/out of the controlled volume based on the divergence theorem and cross-sectional flux method in Varon et al. (2018). It is particularly useful in basin-wise methane emission calculation and is commonly used for in-situ aircraft measurement of gas plumes (Cambaliza et al., 2013). Similar methods were adopted for methane emission calculation by Zhang et al. (2020), Sadavarte et al. (2021) and Schneising et al. (2020).

The column-averaged dry air mixing ratio (concentration) enhancement of methane (ΔXCH_4, in ppb) is first calculated by subtracting a background methane mixing ratio XCH_{4_base} from the retrieved atmospheric XCH_4. For monthly emission estimation, $XCH4_{\_base}$ is taken as the 10th percentile of the monthly aggregated XCH_4 in the domain (de Gouw et al., 2020). Our analysis shows that the 10th percentile is consistent with the average XCH_4 for a nearby region free of suspected methane sources. For daily emission estimation calculations, the 10th percentile approach is not feasible due to limited data coverage and results in random unexplainable fluctuations in daily XCH_{4_base}'. In this case, the moving average method is adopted. An average of XCH_4 for a selected nearby region free of suspected methane sources over a 15-day period leading up to the date of evaluation is used as XCH_{4_base} in daily emission estimation. This method provides a more timely estimation than the one for monthly background and ensures the progressive transition of the daily background estimation. The monthly methane background is compared with the daily methane background in Figure 2 for the Permian Basin during 2019/2020. The background level of the daily calculations aligns with the trend corresponding to the monthly calculations, demonstrating micro fluctuations and ongoing variations within the course of a month.

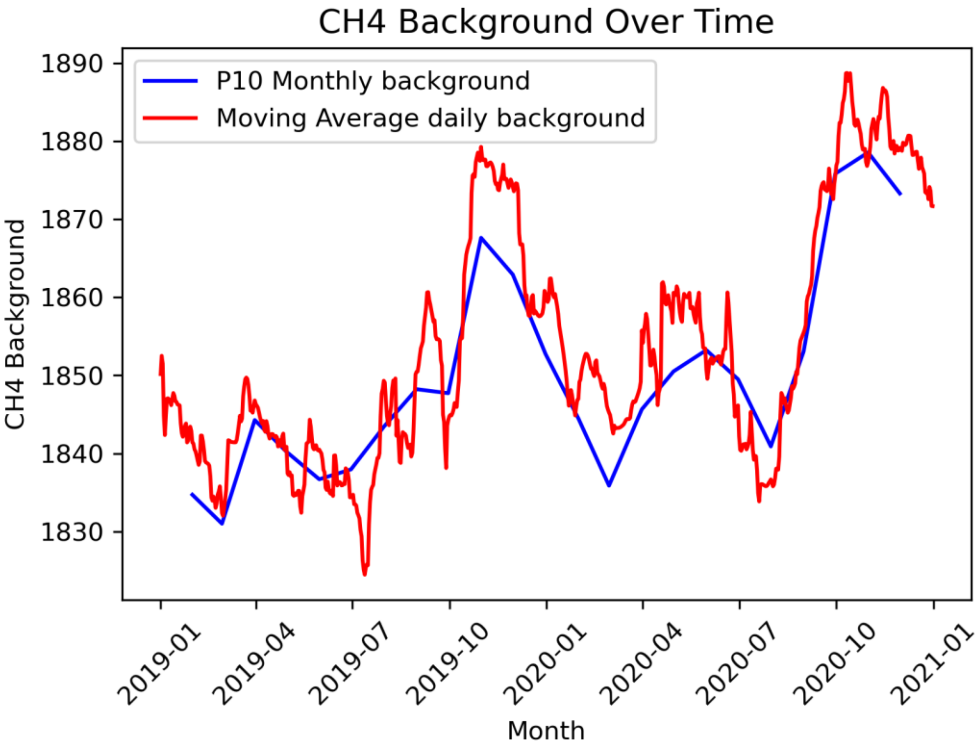


Figure 2. Blue: 2019-2020 Permian Basin background methane mole fraction in ppb for monthly methane estimation using the 10th percentile approach. Red: 2019-2020 Permian Basin background methane mole fraction in ppb for daily methane estimation using a moving average with a 15-day window.

For validation and application purposes, the retrieved column-averaged methane mixing ratio XCH_4 from S5p is, in fact, a representation of the product of dimensionless averaging kernel A_{CH4} and the true column-averaged methane enhancement XCH_4 divided by the dry air column

$V_{\mathrm{air,dry}}$ (in the unit of m^{-2}) calculated from the surface pressure and water vapor profile (Hasekamp et al., 2019) which are outputs of S5p products. Averaging kernel A_{CH4} acts as a scaling factor to the column-averaged methane mixing ratio XCH_4 for a more accurate representation of the methane concentration at different surface altitudes. Therefore, the column-averaged mass enhancement of methane over the background level $\Delta\,\Omega$ in $\mathrm{g/m}^2$ can be calculated as:

$$\Delta\,\Omega\ = \frac{M_{CH_4} \cdot \rho_{air,dry}}{\mathrm{A}_{\mathrm{CH}_4}} \Delta XCH_4. \qquad \text{Eq. 1}$$

M_{CH_4} is the molar mass of methane in units of g/mol, $\rho_{air,dry}$ is the mean dry air column in $\mathrm{mol/m}^2$. $\mathrm{A}_{\mathrm{CH}_4}$ is taken as the near-surface averaging kernel in this application. $\Delta\,\Omega$ is the mass load of methane added to a volume of the atmosphere by anthropogenic activities that are believed to be above and beyond what would be added by natural/background emissions.

The methane flux Φ in the direction of the wind speed must be equal to the product of $\Delta\Omega$, the wind speed v, and the length of the surface Δl perpendicular to the wind. However, the direction of the windspeed often varies spatially and temporally. It is not computationally friendly to estimate methane flux along the direction of the wind with respect to each emission source within the region of investigation. Schneising et al. (2020) solved this problem by coordinate rotation, for which the geographical longitude and latitude of the investigated area are transformed into rotated coordinates so that the zonal direction matches the average wind direction in the area. A workflow is proposed to account for every grid cell's wind direction by decomposing the wind flux into longitudinal and lateral components while keeping the native geographical coordinate system unchanged. The total emission rate can be obtained by summing up all lateral and longitudinal methane flux leaving the boundary of the investigated region. Figure 3A illustrates an example of a wind vector field over a controlled domain, and Figure 3B demonstrates the decomposition of fluxes along the direction leaving the controlled domain. To honour mass balance within the region, the assumptions are: 1) Methane flux entering the controlled volume is negligible – It is assumed that no methane sources are present outside the controlled volume; hence, the methane enhancement from outside is zero. 2) No methane accumulation within the controlled volume – All methane emitted within the region leaves the controlled volume through the boundaries. The wind data used in this study were collected from a recognized database: ERA5-Land reanalysis dataset (Muñoz Sabater, 2019). The spatial discretization of the wind data are down scaled to match the spatial resolution of the methane data. The 10-meter height wind speed was used in the calculation to match up with the near surface methane concentration. While through an S5p processor version update in April 2019, the horizontal and vertical components of the wind at 10-meter height data were added to the S5p level 2 support data package (Landgraf et al., 2021) to support the transport analysis at the surface. The effective wind speed used in the calculation can be approximated using the wind data products provided in the package.

Then, the methane flux $\varPhi$ leaving the domain V can be related to column-averaged mass enhancement $\varDelta\ \varOmega$ by:

$$\varPhi(V) = \int \varDelta\varOmega_b \cdot v_b\, dl. \qquad \text{Eq. 2}$$

$\varDelta\varOmega_b$ is the column mass enhancement at the boundary, v_b is the wind component perpendicular to the boundary, and l is the length of the boundary. More clearly, the boundary refers to the outer boundary of the larger investigated domain V for the calculation rather than individual grid cell. Since multiple sources are present in V, and the methane column enhancement in V does not reflect immediately at the boundary, $\varDelta\varOmega_b$ at the boundary is estimated using all $\varDelta\varOmega_n$ from individual V_n in V along the direction of the flux leaving the boundary. V_n denotes the discrete volumes composing the aggregate volume V.v_b at the boundary is equal to the mean magnitude of the wind components along the direction of the flux leaving the boundary. Therefore, the total methane flux $\varPhi(V)$ leaving the boundary of domain V can be obtained with the discrete summation of all the boundary flux $\varPhi_b$ leaving the domain V (outflow flux):

$$\varDelta\varOmega_b = \frac{\sum_1^n \varDelta\varOmega_n}{n}; \qquad \text{Eq. 3}$$

$$v_b = \frac{\sum_1^n v_n}{n}; \qquad \text{Eq. 4}$$

$$\Phi(V) = \sum_{1}^{m} \Phi_b = \sum_{1}^{m} \Delta\Omega_b \cdot v_b \cdot \Delta l_b \,. \qquad \text{Eq. 5}$$

n is the number of v_n orthogonal to Δl_b and enclosed by the dashed-border rectangle in Figure 3B, and m is the total number of boundary flux Φ_b leaving V. In Figure 3C, $\Delta\Omega_{b1}$ and v_{b1} are the average of $\Delta\Omega_1$, $\Delta\Omega_2$ and v_1 and v_2 respectively.

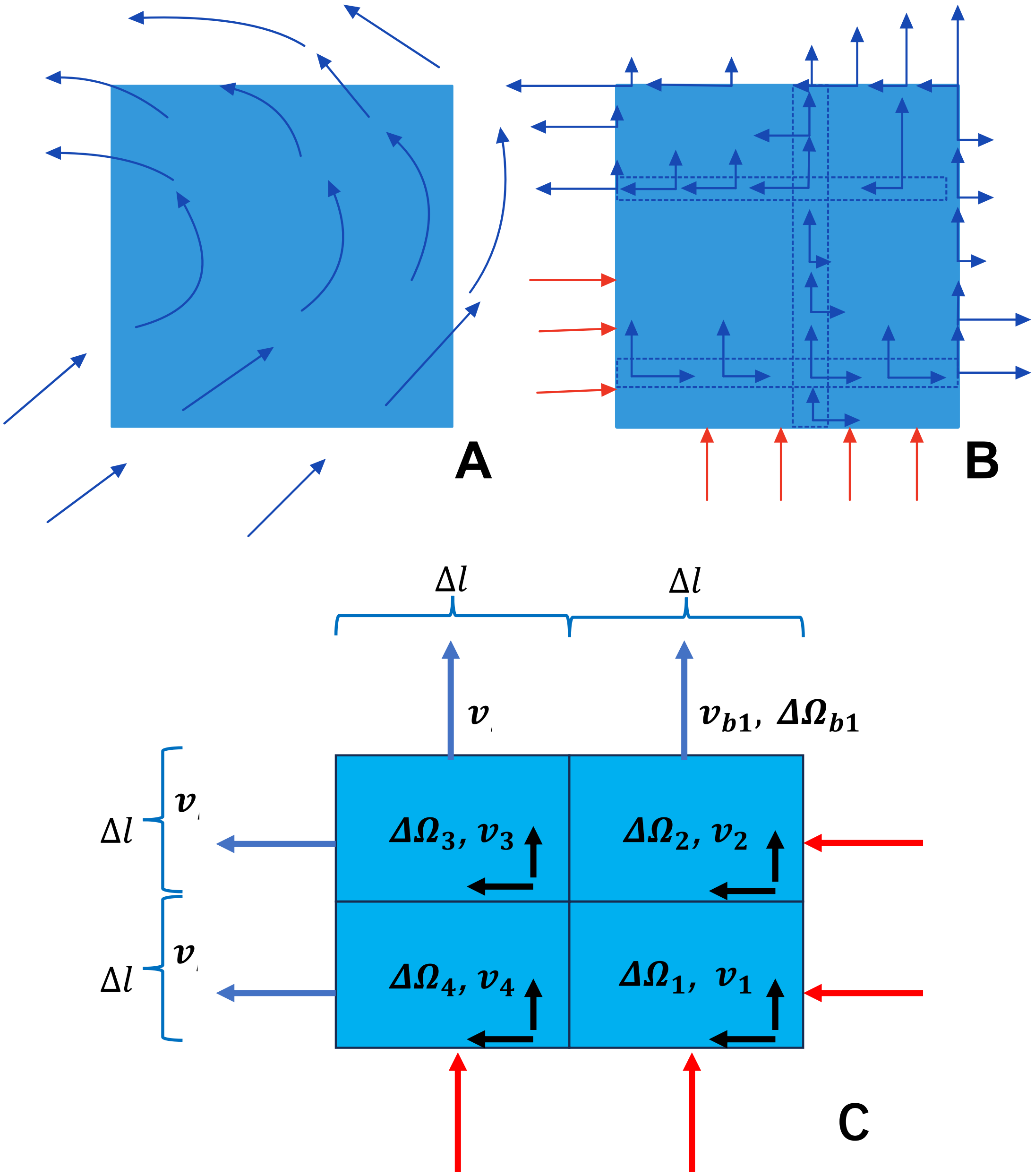


Figure 3. A): Example of a wind vector field over a defined domain V. B): Demonstration of methane flux decomposed into longitudinal and lateral flux. Blue arrows at the boundary represent outflow flux, while red

arrows represent inflow flux assumed to be zero. The outflow flux at the boundary is estimated using the components of all the flux perpendicular to the outflow face (an example is shown with the dashed-border rectangle). C): Demonstration of a sample domain V consists of 4 grid cells.

3.2 Results

This chapter investigates two major U.S. hydrocarbon production regions – the Permian Basin and the Appalachian Basin. Both of these regions have unique characteristics regarding sources of methane emissions, with the Permian Basin being the largest oil-producing basin in the country and the Appalachian Basin being a complex blend of shale gas production, coal mining industry emissions, and proximity to urban emission sources. The presence of multiple sources of emissions in these regions can introduce substantial uncertainties in the analysis of methane emission rates. Examining these basins offers valuable insights into the challenges of accurately estimating methane emissions in large hydrocarbon production regions, characterizing emission sources, and informing the development of more effective mitigation strategies. In this study, we focus our analysis using the monthly methane emission estimation rather than the daily data. The results presented here are based on the analysis of monthly averages only.

3.2.1 The Permian Basin

There have been nearly 500,000 registered wells drilled or to be drilled in the highly prolific Permian Basin since its initial development in the early 90s. As of November 2022, there are nearly 163,000 active wells, with oil wells comprising 70% of the total. Furthermore, there are thousands of completed wells with unproduced oil and gas ready to be produced based on data extracted from Enverus PRISM. As a result, as reported by EDF, the Permian Basin may be the largest emitter of methane of all the oil and gas plays in the United States. In many instances, natural gas, considered less valuable than liquids, is often flared or vented out directly into the atmosphere, as the level of production activities far outpaces the construction of gas transportation pipelines. Flaring and venting, along with fugitive methane leakage from facilities, pipelines, and wellheads, are the main methane emission sources. Yu et al. (2022) also demonstrated with a few aerial campaigns that the methane emission from gathering pipelines in the Permian Basin is 14 – 52 times higher than the EPA's estimate and 4 – 13 times higher than the highest estimate derived from bottom-up measurements of gathering pipelines.

The investigated area for this study includes the Permian Basin and surrounding regions in western Texas and southeastern New Mexico of the United States (30° to 34.5° N and 100° to 105° W). The illustration of the interpolated monthly methane enhancement for the area of investigation is presented in Figure 4. The estimated monthly methane emission rates for the period January 2019 to February 2022 are shown in Figure 6. The aggregated mean methane enhancement over the investigated period is shown in Figure 5. It is noticed that methane emissions are persistently high in the northern and eastern parts of the Delaware basin, where unconventional shale gas has been actively explored and developed in recent years. Central Midland Basin, where a lot of the new horizontal wells were drilled and produced oil and gas. Over 22,000 flaring events were captured in the Permian Basin in 2019; over half are found in the Delaware Basin (Data from Enverus PRISM). The analysis of the background methane mole fraction during the investigation period reveals a consistent annual trend with strong seasonality characteristics. The study found that the background methane mole fraction exhibits two peaks each year, occurring around April and October. These seasonal fluctuations are likely linked to the seasonal availability of methane-producing sources such as wetlands and agriculture, which have been identified as the top two sources of methane emissions by the International Energy Agency (2022). In the study of methane concentration in North America, Javadinejad et al. (2019) found that an increase in methane levels has a strong correlation with low vegetation coverage and high temperatures, providing insight into the underlying mechanisms driving the seasonal variations in the methane cycle that are observed in this study.

The mean methane emission estimated from the study domain in this period is 3.56 Mt/year, with annual emissions of 4.46 Mt/year for 2019, 3.42 Mt/year for 2020 and 2.84 Mt/year for 2021. The 2019 estimation is higher than the 2.9 Mt/year proposed by Zhang et al. (2020), but the investigation time frame is different than ours (May 2018 to March 2019 by Zhang et al. vs the full year of 2019 in this study). Schneising et al. (2020) also estimated an average methane emission rate of 3.16 Mt/year for 2018/2019, which is also smaller than our estimation for 2019. These minor discrepancies can be attributed to differences in the study period, study domain coverage, and the utilization of TROPOMI measurements. We have utilized all the high-quality measurements within the time frame, while Schneising et al. (2020) utilized a selective subset of daily measurements with high data spatial coverage but at the expense of low temporal coverage.

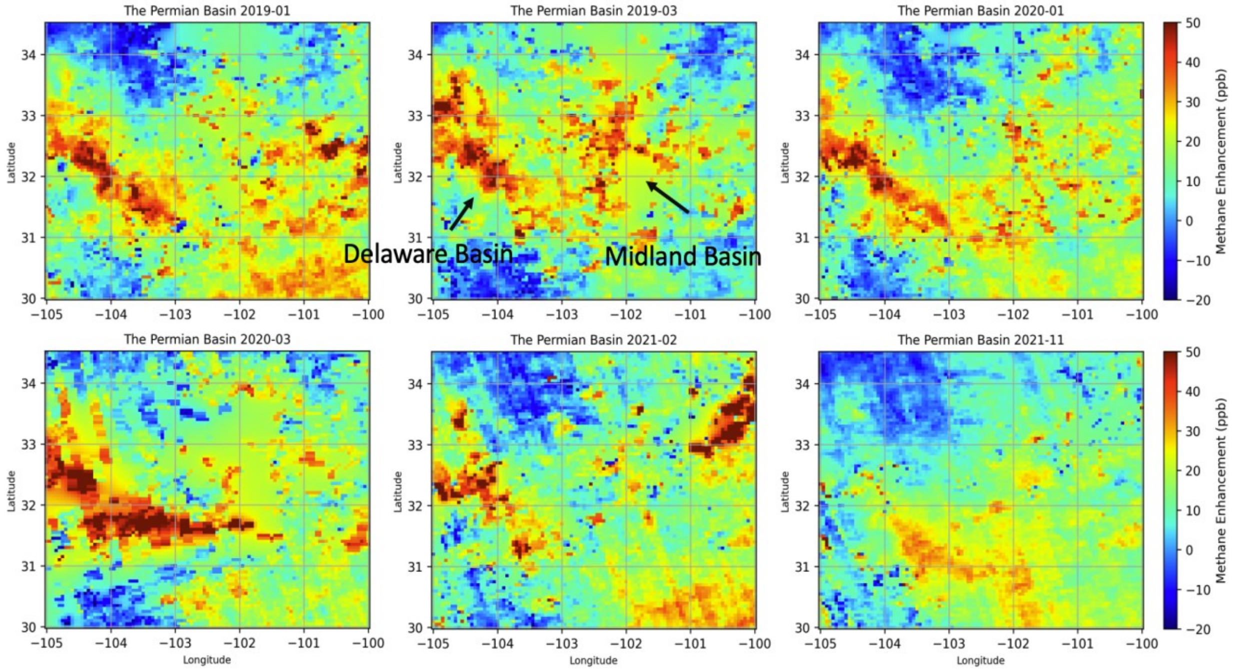

The Permian Basin 2019-01
The Permian Basin 2019-03
The Permian Basin 2020-01
The Permian Basin 2020-03
The Permian Basin 2021-02
The Permian Basin 2021-11
Latitude
Longitude
Delaware Basin
Midland Basin
Methane Enhancement (ppb)
50
40
30
20
10
0
−10
−20
34
33
32
31
30
−105
−104
−103
−102
−101
−100

Figure 4. Examples of monthly interpolated methane enhancement map of the investigated area for the Permian Basin. The monthly methane enhancement is obtained by averaging daily retrievals within the month and then filling in the missing measurements using IDW.

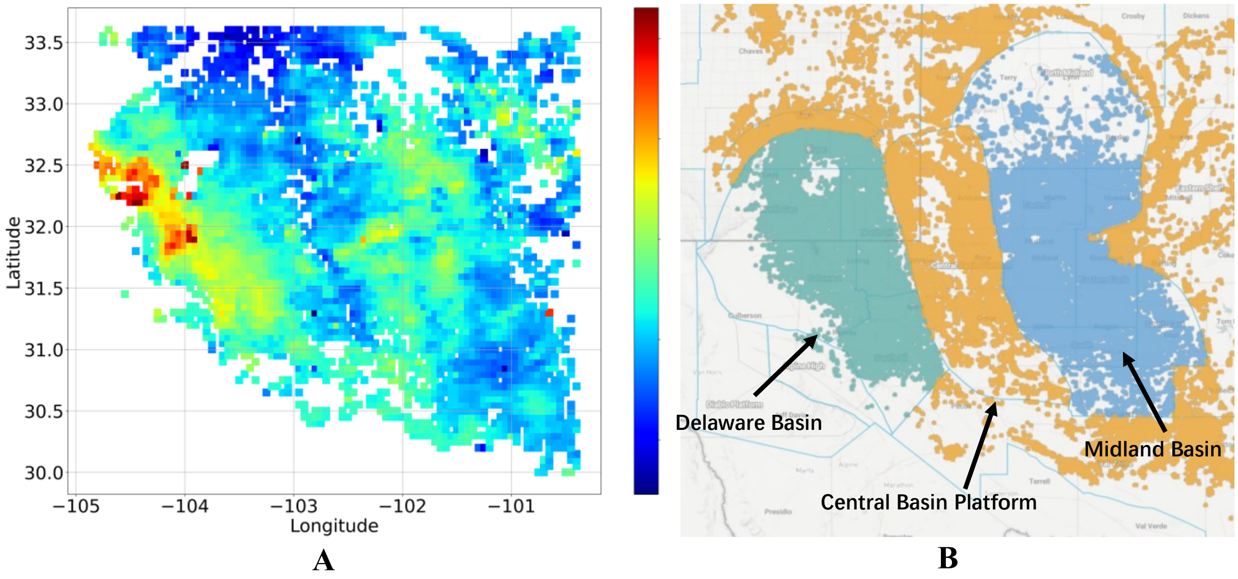


Figure 5 A):2019-2021 mean methane enhancement map for the Permian Basin. Only measurements of grid cells with producing wells are plotted in this figure. B); Operating wells in the Permian Basin colored by their operating sub-basin (Mapped using Enverus PRISM).

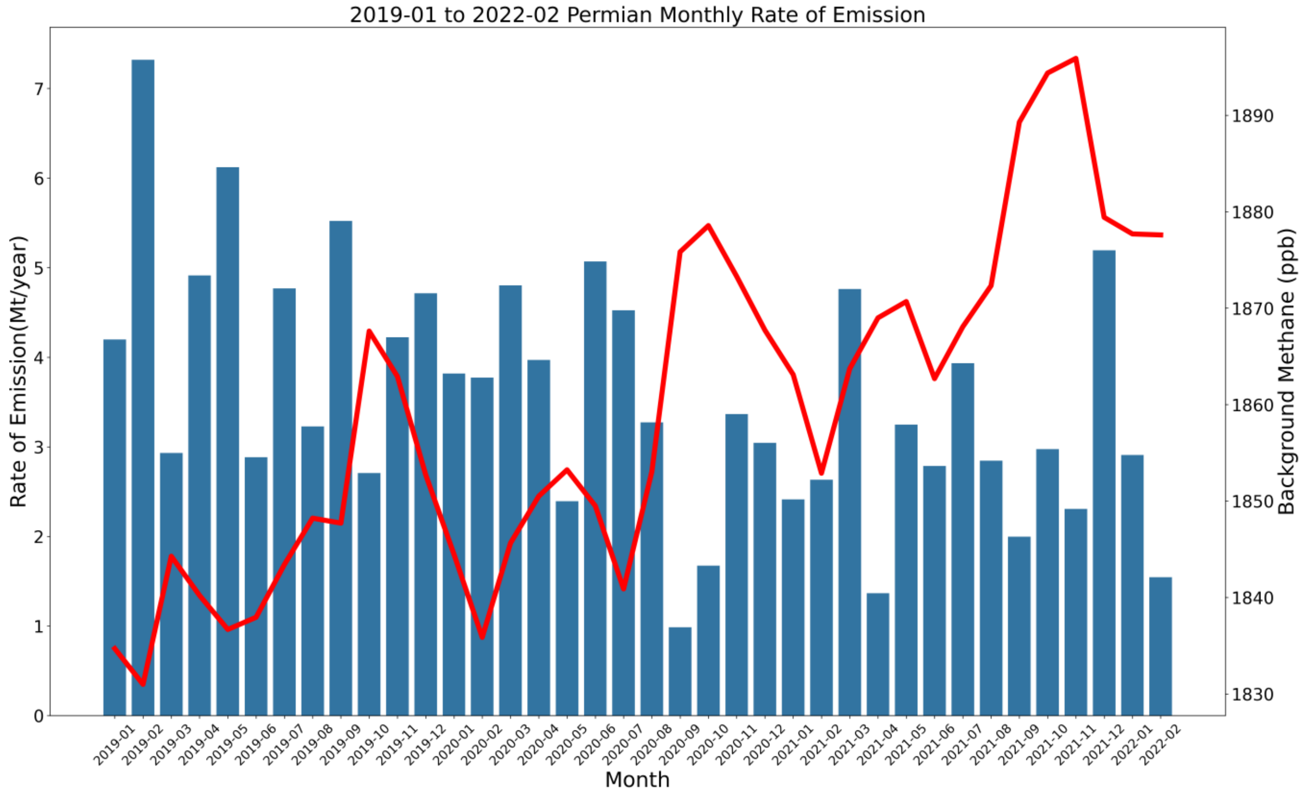


Figure 6. Monthly methane emission rate for the target domain encompassing the Permian Basin. Blue bars indicate the rate of emission, and the red line denotes the corresponding background methaneconcentration XCH_{4_base}.

3.2.2 The Appalachian Basin

Due to the discovery of a large quantity of gas resources in the Devonian Marcellus Shale and Upper Ordovician Utica Shale, along with being the country's leading producing region of coal, the Appalachian basin emits more methane than the Permian Basin, according to our study. As discussed in the Permian Basin case, other sources, in addition to shale gas production, have contributed to the overall emissions. Our study shows that strong methane enhancement is frequently observed around regions with mining operations (especially in southwest Pennsylvania and northern West Virginia) (Figure 7). According to the 2020 U.S. Coalbed Methane Outreach Program from EPA, about 61% of methane emissions from mining activities are air ventilated from underground mines containing low methane concentrations. The other 39% are from abandoned coal mines, surface mining, post-mining operations and degasification systems at underground coal mines. Major cities and farmlands can also be found within the investigated region in Ohio, where strong methane enhancement is observed. Therefore, coal mining is inferred to be the major source of methane emission in the Appalachian Basin. Urban and agricultural methane emissions are also contributors in the Appalachian Basin.

Due to the elongated shape of the Appalachian Basin, it is divided into two regions: the southwest region (37.75° to 41.25° N and 79° to 82.75° W) and the northeast region (41.25° to 42.4° N and 75.4° to 78.9° W). We have estimated that nearly 65% of the Appalachian Basin's methane emissions originate from the southwest region. The combined mean methane emission of both areas from 2019 to 2021 using the monthly estimation amounts to 4.46 Mt/year with annual emissions of 4.92 Mt/year for 2019, 4.32 Mt/year (2020) and 4.14 Mt/year (2021). Results for the monthly emission rate for the southwest region are shown in Figure 8. Schneising et al. (2020) also estimated an average methane emission rate of 2.36 Mt/year for the period of 2018/2019 – approximately 50% of our estimation for 2019. Schneising focused on two specific hotspot regions in the southwestern and northeastern parts of Pennsylvania, while we consider the entire stretch of the Appalachian Basin. Furthermore, it should be noted that only 24 days contribute to this 2-year emission estimation calculations by Schneising et al. (2020) due to the rigorous data filtering approach and high cloud coverage in the area. We implemented a different strategy by keeping all high-quality data so that all the confident measurements are used to the greatest extent.

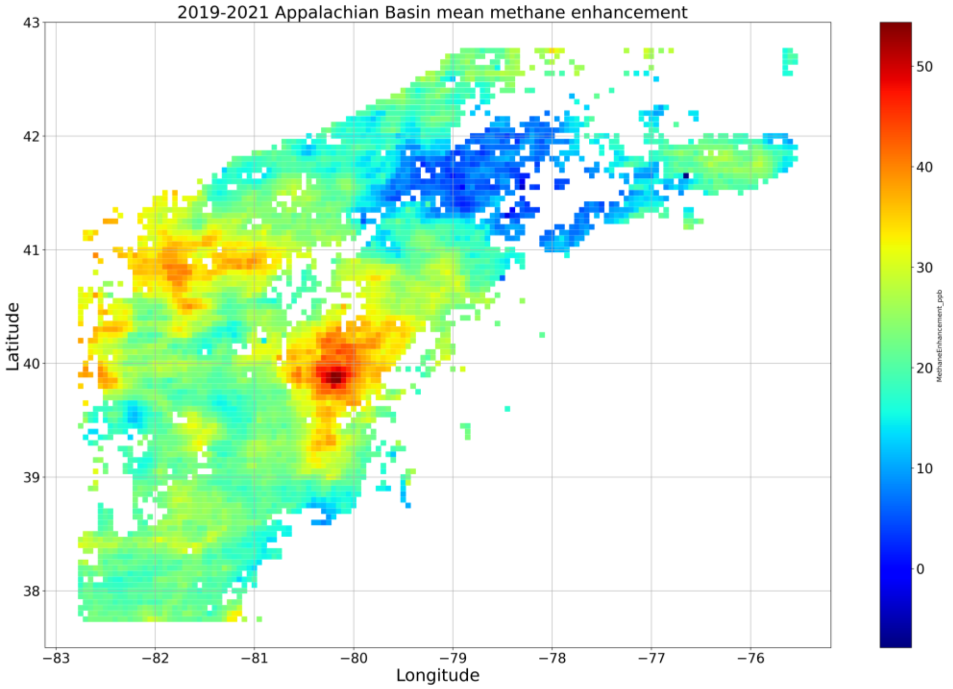


Figure 7. 2019-2021 Mean methane enhancement map for the Appalachian Basin. Only measurements of grid cells with producing wells are plotted in this figure.

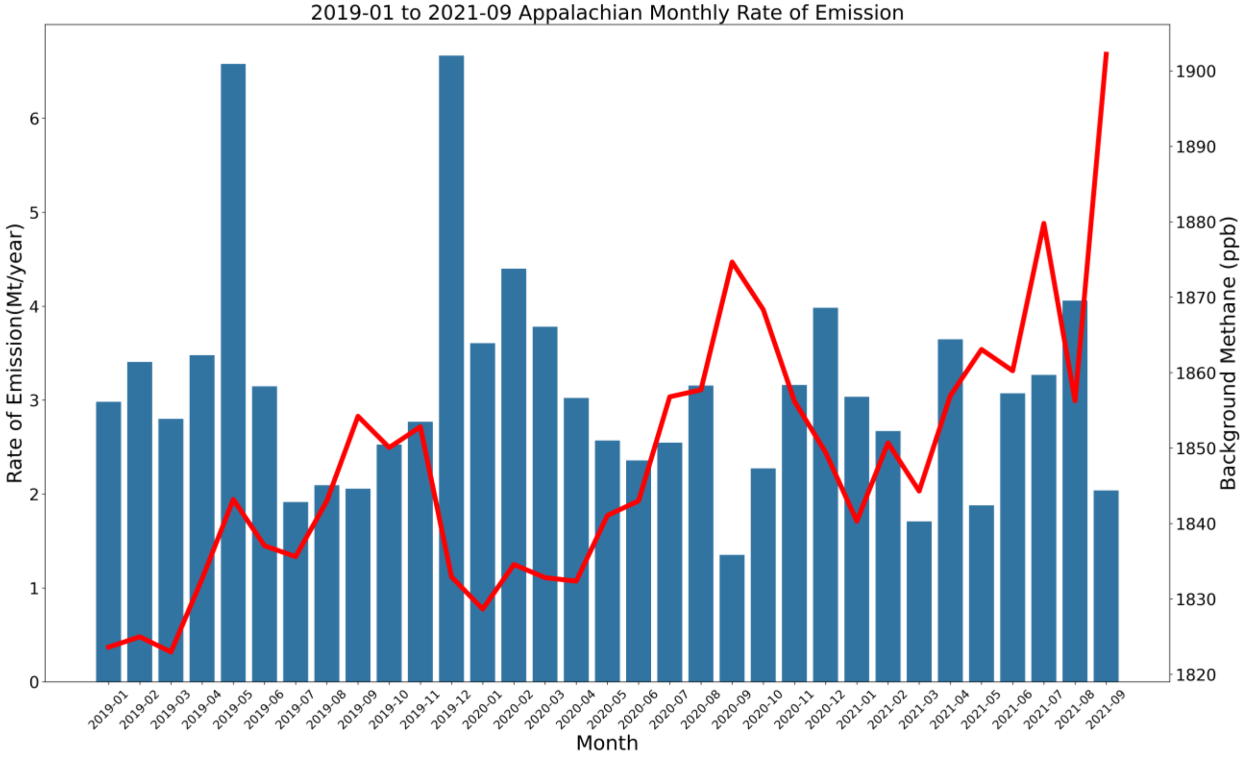


Figure 8. Monthly methane emission rate for the Southwest region in the Appalachian Basin. Blue bars indicate the rate of emission, and the red line denotes the corresponding background methane mixing ratio XCH_{4_base}.

3.3 Uncertainties

The estimation of methane emission rates is subject to several potential sources of errors and uncertainties, including the assessment of the background methane mole fraction, effective wind speed estimation, raw data quality, aggregation of TROPOMI observations, interpolation to fill in the missing data, and assumptions made in the mass balance method.

The background methane mole fraction is a critical variable for accurately determining the level of methane enhancement above the normal levels of methane originating from within the region. The strong seasonality observed in the background methane concentration adds another layer of complexity to its estimation. Several approaches for estimating the background methane mole fraction are investigated. The analysis from the Permian Basin shows that usi^{ng} the 10th percentile of the XCH_4 values within the area of investigation is consistent with the average XCH_4 of an upwind region without noticeable methane emissions. This simple criterion provides a valid estimation of the meteorological conditions.

The uncertainties associated with estimated methane emission rates are closely dependent on the variation in data quality and the scarcity of daily XCH_4 observations. The study on the Permian Basin found that approximately 80% of the observation days contained more than 50% missing data after quality filtering. Averaging across the 38 aggregated monthly XCH4 observations, the proportion of missing data is approximately 18%, with a peak of 52% occurring in a single month. Higher data sparsity implies a greater need for interpolation, leading to increased uncertainties in the estimated methane emission rates. In addition, it is important to consider the potential impact of different spatial and temporal interpolation methods, as they may lead to slight variations in the estimated values.

Effective wind speed plays a significant role in the estimated emission rate, as it affects the transport of emitted methane out of the region. A key assumption is that all emitted methane leaves the region through its boundary. Still, it is challenging to determine the effective daily wind speed in carrying the methane out of the region precisely, and the result would strongly affect the flux calculations. To minimize uncertainties stemming from varying wind speeds in the estimation, the effective wind speed in this study is determined as the nearest measurement to the time of retrieval for the $XCH4$ data.

A key assumption of this study is that there are no significant methane emission sources outside the region (i.e., there is no additional methane inflow above the background level into the controlled region). This assumption generally holds for isolated (small) controlled regions, for which the mass balance approach was originally proposed. In previous studies, small domains with a few isolated sources were selected, and the coordinates were rotated to be orthogonal to the outflow direction. This assumption may not always be applicable in larger domains, such as an entire producing basin, with many emission sources and neighbouring human activities, including farmlands, small towns, and major cities, that are known to be major emissions sources. Ignoring any methane emission sources outside the controlled region has likely resulted in overestimating the emission rates. In the context of emission monitoring and risk mitigation, one may argue that it is preferable to provide conservative (higher) estimates instead of overly optimistic (lower) ones in situations with high levels of uncertainty or ambiguity. These estimations using satellite measurements and other top-down methods (Alvarez et al., 2018) are always higher than the U.S. EPA inventory estimate, which can be attributed to the existing inventory methods failing to account for emissions released during unexpected conditions (fugitive emissions).

It is also important to acknowledge that besides the uncertainties above, satellite measurements often involve significant uncertainties. Dubey et al. (2023) described these uncertainties as the satellite's minimum detectable limit (MDL), which can vary depending on the source emissions and the atmospheric conditions.

CHAPTER 4: A DATA ANALYTICS APPROACH FOR UNRAVELLING THE COMPLEXITY OF METHANE EMISSIONS

4.1 Introduction

According to the Environmental Protection Agency (EPA) in the United States, the oil and gas industry is the largest industrial source of methane emissions, accounting for nearly 30% of total methane emissions in the country (EPA, 2019). With leaks and venting occurring at every production stage, from drilling and extraction to processing, transportation, and distribution, methane can escape from wells, pipelines, and other equipment, as well as from flaring and venting operations. Despite these challenges, this sector has substantial mitigation and emission reduction potential. The International Energy Agency estimates that global methane emissions from the oil and gas sector could be reduced by up to 75% using existing technologies and best practices (IEA, 2017). Furthermore, there is a growing commitment from numerous companies and governments to curb their methane emissions (World Bank, n.d.; Methane Guiding Principles, n.d.; IEA, 2019). These efforts include, but are not limited to, enhancing monitoring and detection measures, addressing and repairing leaks, phasing out aging equipment, and minimizing flaring and venting operations.

Multiple studies have been conducted to evaluate the ability of the Sentinel-5 Precursor (S5p) satellite and its Tropospheric Monitoring Instrument (TROPOMI) measured products to estimate methane emissions on a regional scale (Bian et al., 2023; Varon et al., 2018, 2019; Lyon et al., 2021; de Gouw et al., 2020; Zhang et al., 2020; Sadavarte et al., 2021; Schneising et al., 2020, Shen et al., 2022), and these studies have consistently shown the significant potential of the S5p for this application. Given these findings, it is apparent that the methane concentration measurements obtained from S5p have the potential to extend their utility beyond estimating regional emissions. It contains valuable insights that can potentially improve our understanding of methane emissions within the oil and gas industry. While some studies have examined the correlation between satellite-measured methane enhancement and a few oil and gas operating parameters to some extent, these investigations are not very comprehensive. Regional or geological factors, temporal dynamics including seasonal variations and influences from various midstream facilities have not been adequately explored. Long-term trends and changes in methane enhancement and its correlation with additional oil and gas operating parameters also require further investigation with alternative statistical or machine learning methods. Many studies investigate based on the reported methane emission sources and intensities, but the reported numbers often lack adequate characterization of sources due to discontinuous emission events from P&A or orphaned wells and other fugitive leakage from Superemitters. Moreover, underestimation is always found in the reported emission numbers when compared with results from the studies mentioned above using top-down estimation methods.

This chapter aims to investigate the relationship between 2019-2021 Permian Basin methane concentration measurements obtained from S5P and various oil and gas operating parameters extracted from Enverus PRISM more comprehensively using advanced data analytics techniques. This analysis contributes to a growing body of knowledge on the factors influencing methane emissions in the industry. The research was designed to provide insights into the conditions under which methane emissions occur. By combining different data analytic techniques, we aimed to understand the heterogeneous nature of methane emissions across different operational and geographical contexts.

4.2 Methods

4.2.1 Dataset

The same Level 3 processed monthly methane enhancement data with spatial resolution 0.05° latitude × 0.05° longitude methane as how it was . However, correlating methane enhancement at a resolution of 0.05° by 0.05° presents significant challenges due to inconsistencies in the spatial and temporal scales between methane enhancement data and the other correlating features. These discrepancies can lead to uncertainties in analyzing the relationships between methane emissions and oil and gas operational parameters. To address this scale discrepancy, we implement an upscaling approach for the correlating features, enabling them to match the resolution of the methane enhancement data. Upscaling is the process of reducing the resolution of a dataset by aggregating finer-scale data into coarser-scale representations (Blöschl et al., 1995), which facilitates comparisons with the methane enhancement data at larger spatial scales. For instance, we identify all wells situated within the geographical extent of each grid cell and aggregate their relevant attributes, such as the total number of wells, total gas produced, oil produced, operator information, well status, and other correlating features, at the grid cell level. This spatial aggregation process ensures that the oil and gas correlating features are consistently represented at the same spatial resolution as the methane enhancement data, enabling a more accurate and reliable examination of the interrelationships between these variables. Temporal discrepancies between daily methane measurements from S5p and oil and gas correlating features are also addressed. We adopt a temporal aggregation approach, transforming daily methane measurements into monthly averages to align with the temporal resolution of the oil and gas correlating features.

4.2.2 Workflow

Initially, Random Forest regression and classification models, implemented in Python's scikit-learn package (Pedregosa et al., 2011), were deployed to identify the key features correlating with methane enhancements. We acknowledged the limitations of these models due to the inherently variable nature of methane emissions, prompting us to incorporate an additional unsupervised learning technique – clustering analysis. The Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm was chosen for the clustering analysis (Mclnnes et al., 2017). The K-means++ algorithm (Arthur et al., 2007) was also applied to partition the dataset into three distinct clusters for comparative purposes. Following the clustering analysis, we applied the Random Forest regression model to each of the two redefined clusters.

4.3 Emission Characteristics and Sources Analysis

4.3.1 Descriptive Analysis

The dataset is characterized by substantial heterogeneity in operational practices – variability in drilling, completion, and production standards adopted by different operators may lead to disparate methane emission levels. There are over a thousand operators in the Permian Basin (Enverus, PRISM), each contributing to the observed methane concentrations in different ways and quantities. Incorporating many operator classes into a regression model may pose considerable challenges due to the high dimensionality and potential multicollinearity among the features. The approach of allocating methane concentration data to different operators based on their geographic locations is adopted. By spatially associating methane emissions with specific operators, we can reduce the complexity of the regression model and potentially identify key contributors to the observed methane concentrations. This approach allows for a more focused analysis of how different operators contribute to methane emissions.

To determine the appropriate number of operator categories or labels, we plot the cumulative market share of operators in terms of the number of wells they owned (Figure 9). The plot reveals that the first 100 operators account for over 90% of the market share, and the top 40 own over 70% of the wells. Therefore, the number of operator classes is assigned as 41, with the remaining 900+ operators described as "Others".

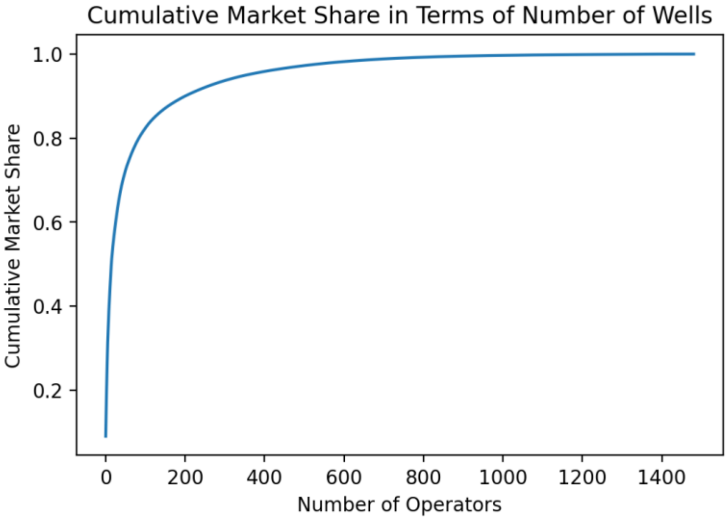
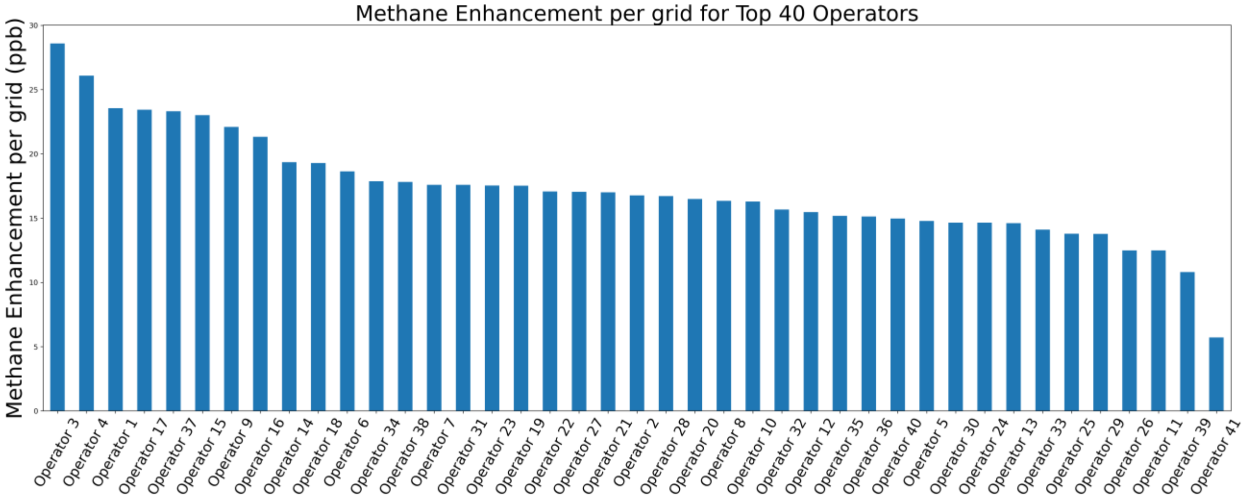
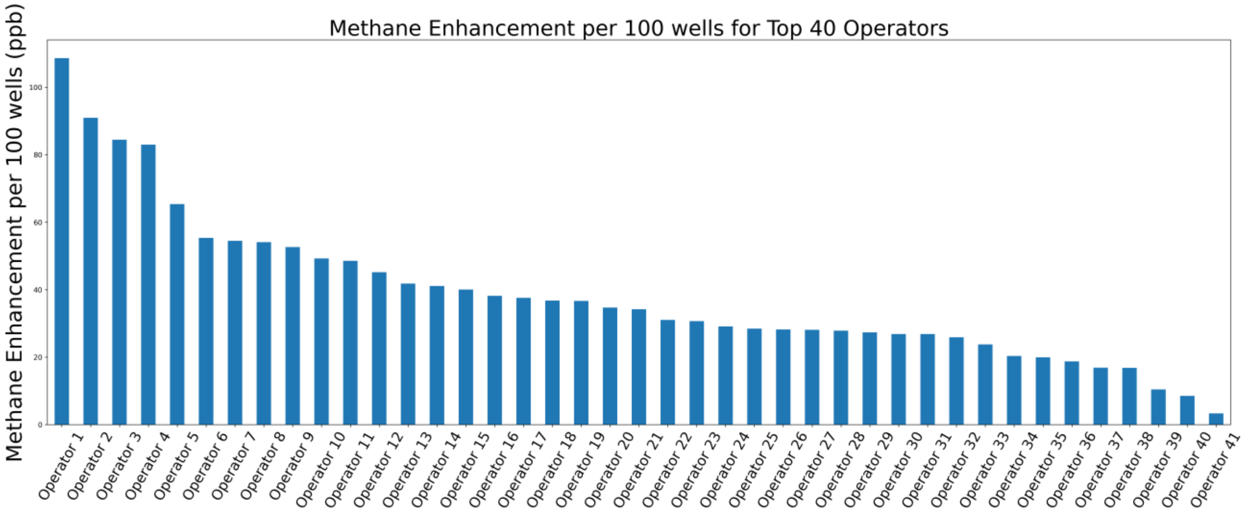


Figure 9. Cumulative market share vs. operator number in descending order.

In this analysis, we utilize level-3 (L3) processed monthly methane enhancement data, expressed in parts per billion (ppb), for each 0.05° by 0.05° grid cell. Assuming that the methane enhancement is exclusively sourced from oil and gas production activities, we allocate the enhancement values to operators operating within each grid cell normalized over 2 factors: well count and production volume, and we also calculate the grid cell average enhancement over the period.

Figure 10 presents the methane enhancement for the top 40 operators in the Permian Basin, ranked based on methane enhancement per grid cell, per 100 wells, and per 1,000 barrels of oil equivalent (BOE) in descending order. It is observed that the variations in methane enhancement among different operators are the smallest in the methane enhancement per grid cell plot, while they are the most diverse in the methane enhancement per BOE plot. This observation supports the hypothesis that there are discernible differences between operators concerning methane emissions. Moreover, certain operators, such as Operators 1, 3, 4, and 9, are consistently ranked within the top 10 across two of the ranking plots. Notably, the majority of their operations are concentrated in the Delaware Basin. This observation highlights the potential regional influence on methane emissions. The substantial disparities observed among operators in the methane emission per BOE plot suggest that production volume may not be the predominant factor influencing methane emissions.





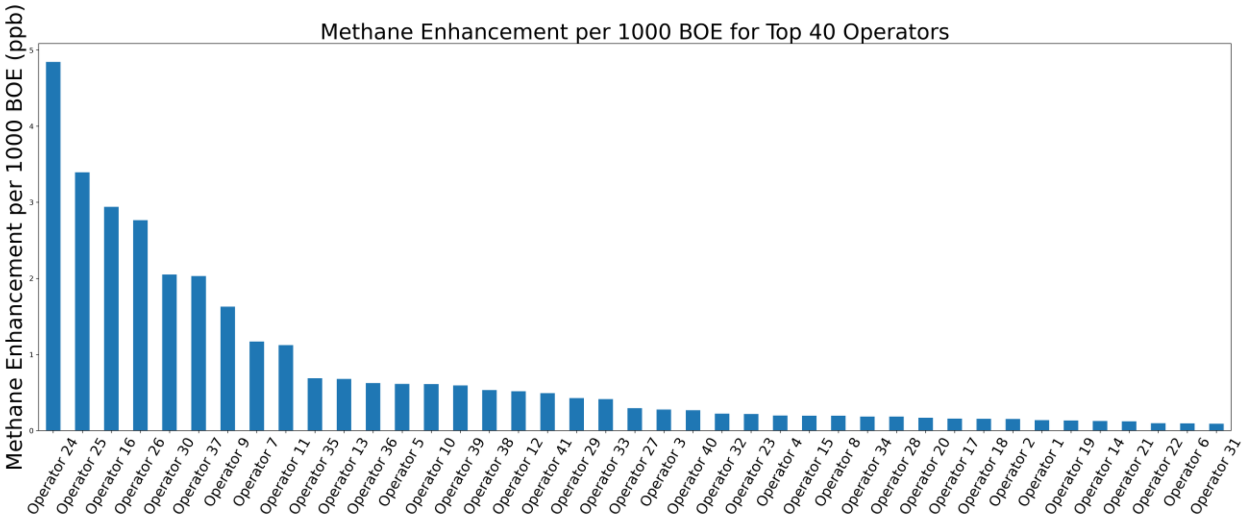
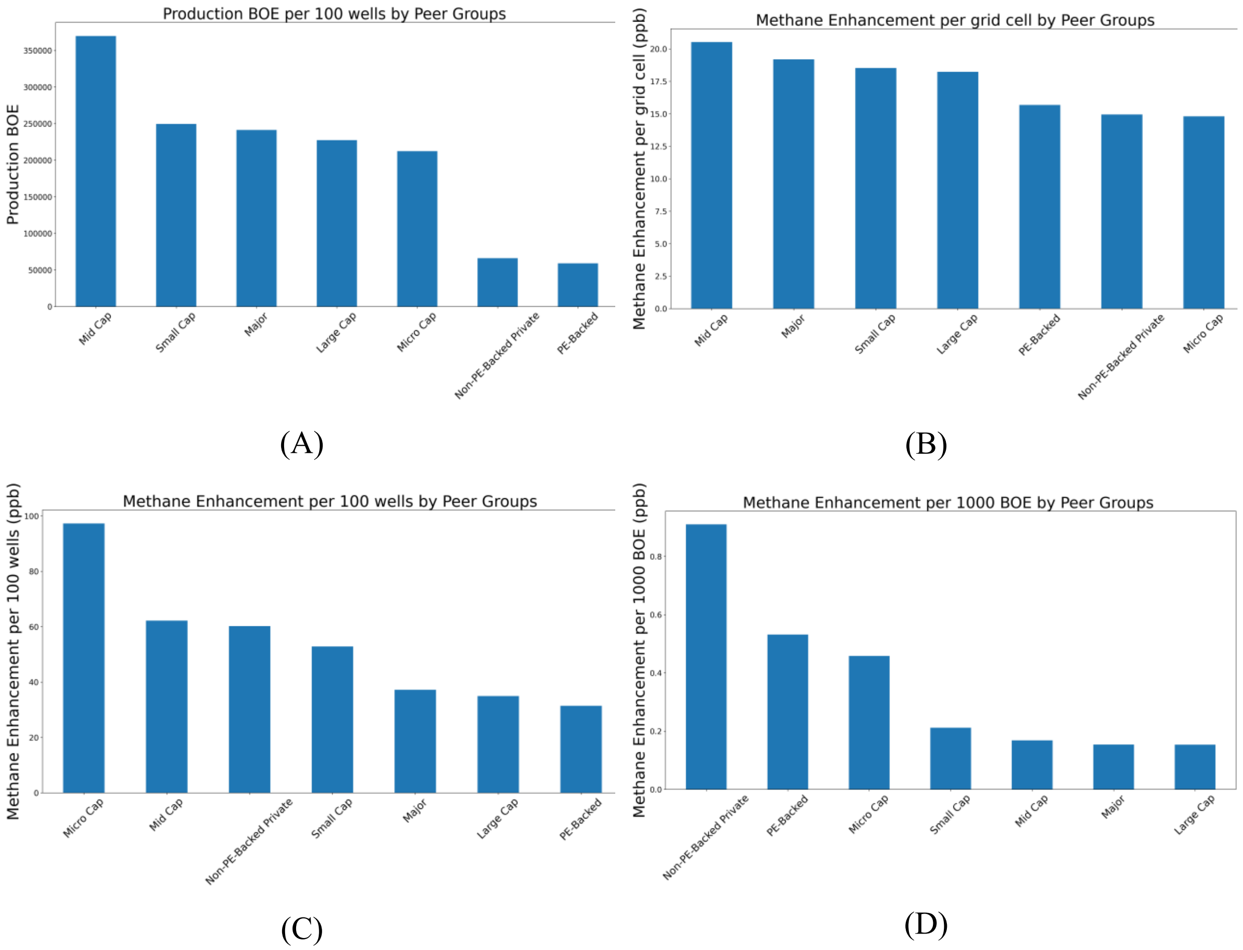


Figure 10. Methane enhancements of the top 40 operators in the Permian Basin: (1) Methane enhancements per grid cell (Top panel), (2) Methane enhancements normalized to every 100 wells (Middle panel), and (3) Methane enhancements per 1000 barrels of oil equivalent (BOE) (Bottom panel).

Next, we examine methane enhancement by primary peer group operating within each grid cell and rank them based on the same three criteria. From the results, private operators typically have lower productivity than their larger capital investment counterparts (Figure 11A), but they generally contribute to higher methane enhancement per barrel of oil equivalent (BOE) production (Figure 11D). In particular, micro-cap operators, having the lowest methane enhancement per grid cell (Figure 11B), are also associated with the highest emissions per 100 wells (Figure 11C): these operators employ lower well density and, hence, can achieve lower overall emissions per grid cell; however, the higher methane emissions per 100 wells indicate that, on a well-by-well basis, their wells are likely emitting more methane. This difference could be attributed to differences in operational practices, technologies employed, emission reduction strategies and other factors.



(A)

(B)

(C)

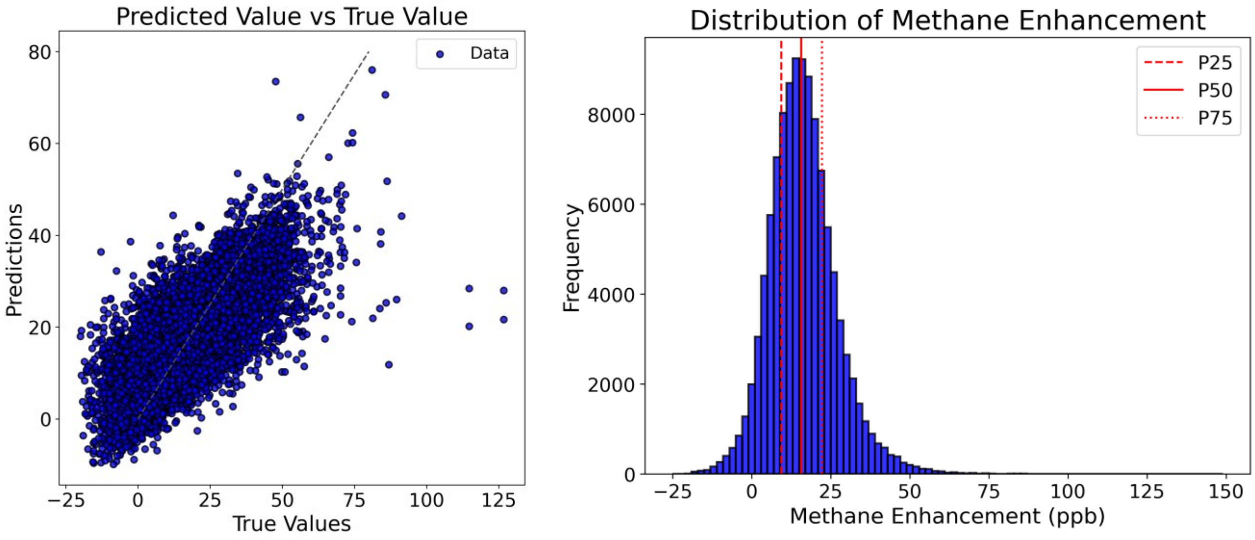
(D)

Figure 11. A): Production volume per 100 wells by operator Peer Groups. B) Methane enhancement per grid cell by Peer Groups. C) Methane enhancement per 100 wells by Peer Groups. D) Methane enhancement per 1000 BOE by Peer Groups.

4.3.2 Feature Importance Analysis and Regression Modeling

A total of 22 features (Table 1) are carefully chosen from an initial set of 47 potential features as inputs to a regression model for predicting methane enhancements of a particular grid cell. Several key considerations, including relevance to the target variable, avoidance of multicollinearity, feature independence, and model interpretability, guide the feature selection process.

A range of statistical and machine learning models, including regression models, Support Vector Machines (SVM), XGBoost, tree-based models, and neural network models, have been evaluated for their efficacy in this paper. For this dataset, the Random Forest model outperforms other models in terms of the proportion of variance explained by the correlating features and the error between the predicted outputs and true values, illustrating its ability to address various challenges, such as nonlinearity, overfitting, and high-dimensional feature spaces. The model's ensemble approach of combining multiple decision trees enables it to capture diverse aspects of the data and enhance its generalization capabilities (Breiman, 2001). Figure 12A shows the relationship between the predicted output and measured values obtained through applying the Random Forest model and the distribution of methane enhancement measurements shown in Figure 12B. When utilizing all observations as input for the model, the resulting coefficient of determination (R^2) is found to be 0.63 and an RMSE of 6.63 ppb.



(A) (B)

Figure 12. A) Predicted methane enhancement VS True methane enhancement plot. B) Distribution of measured methane enhancements in the Permian Basin within 2019-2021.

The model tends to overestimate lower methane enhancement levels (below the 50th percentile) and underestimate higher methane enhancement levels (above the 75th percentile). Negative observations, which constitute approximately 10% of the dataset, arise from how the background methane concentration is fixed at the 10th percentile level within the study area. These negative measured observations often share similar attributes with instances that have negligible emissions or slightly positive methane enhancements, contributing to the model's overestimation at lower levels. On the other hand, observations with high measured enhancements (above 50 ppb) and those obvious outliers often correspond to unforeseen events, such as methane leaks, which the model cannot accurately capture. As a result, these occurrences result in the underestimation of methane enhancements at higher levels. The limitations of the model in accounting for such anomalous events underline the challenges in accurately characterizing and predicting the full range of methane emissions from oil and gas operations.

Figure 13 lists all the features used in the model and their relative impact on the predicted methane enhancement using mean absolute SHAP (SHapley Additive exPlanations) values and MDI (Mean Decrease Impurity). MDI is a feature importance measure calculated by computing the total reduction of impurity or entropy caused by a given feature across all decision trees in a random forest model. Features that cause the most reduction in impurity or entropy are considered the most important (Breiman, 2001). A higher MDI value indicates a stronger association between the feature and the target variable. SHAP values are calculated by computing each feature's contribution to the model's predicted output and averaging over all possible feature subsets (Lundberg and Lee, 2017). The SHAP values provide a more accurate and interpretable measure of feature importance than traditional measures like MDI because they consider the interaction effects between features in addition to their main effects (Breiman, 2001). A few features that exhibit significant importance in the model are discussed here. Windspeed and Month, as shown by their Mean Decrease in Impurity (MDI) rankings (1st and 2nd, respectively) and their SHAP importance rankings (5th and 1st, respectively), are identified as critical correlating features. The seasonality in the monthly methane enhancement and the impacts of wind speed can be observed by plotting the monthly distribution of wind speed (Figure 14A) and the critical quantiles of monthly methane enhancement (Figure 14B). The results highlight the importance of considering both temporal and meteorological factors when examining the relationships between methane emissions and relevant features. The seasonality may be attributed to numerous factors, such as changes in temperature, which in turn may affect equipment efficiency and operational cycles of the facilities, leading to more or less emissions during certain months.

Figure 15A illustrates the spatial distribution of average wind speed over the Permian Basin, showing a clear increase from east to west. Comparing this to the average methane enhancement distribution in Figure 15B, the pattern suggests that higher average methane enhancement tends to be found in areas with lower wind speeds. However, this area also coincides with the active development of the Delaware Basin in the east. This active development could be a contributing factor to the observed relationship between wind speed and methane enhancement.

Although there is no evident relationship between Surface Altitude and methane enhancement or other features, the Random Forest model assigns a notable level of importance to Surface Altitude, suggesting an underlying causal relationship with methane enhancement. For instance, differences in atmospheric pressure, temperature, or wind patterns at different elevations could influence the migration of methane gas emitted. Additionally, the importance of surface altitude might suggest that specific geographic regions, characterized by distinct elevations, have unique operational or geological factors that impact methane emissions from oil and gas activities. In other words, surface altitude potentially functions as a proxy for geographic location.

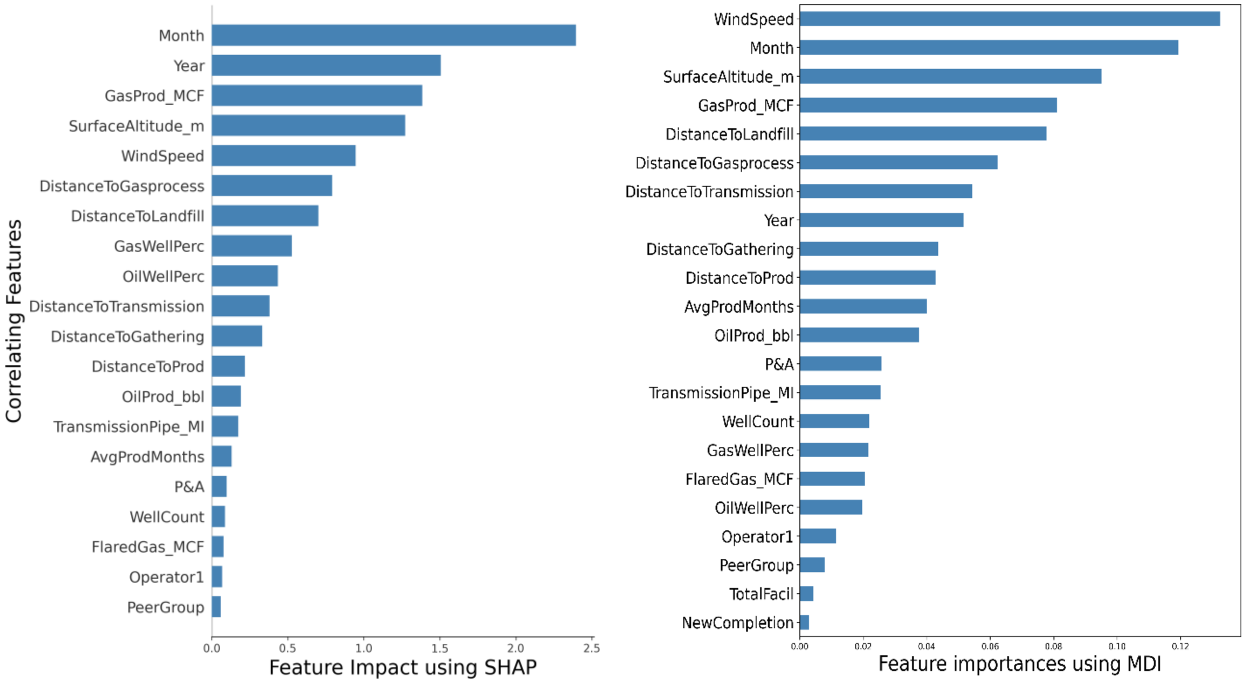
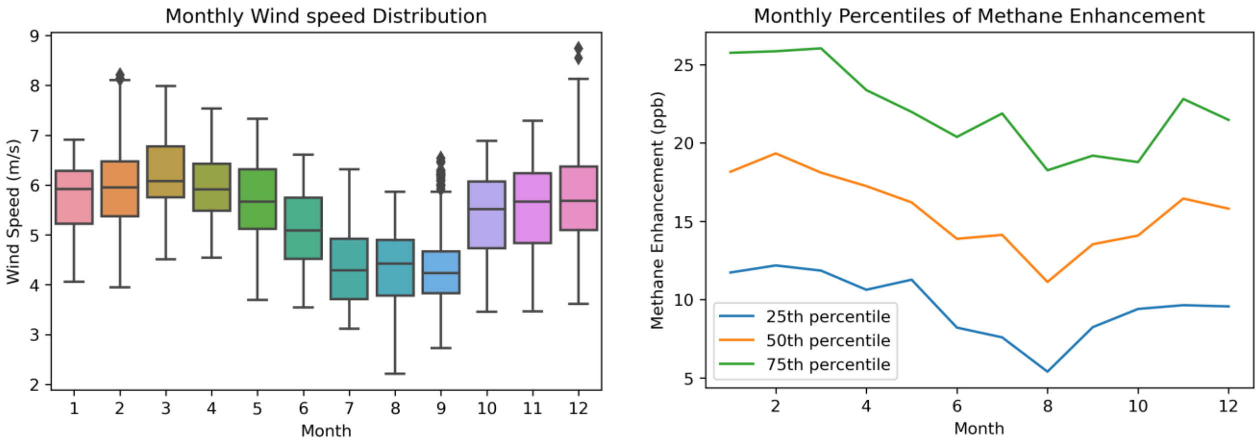


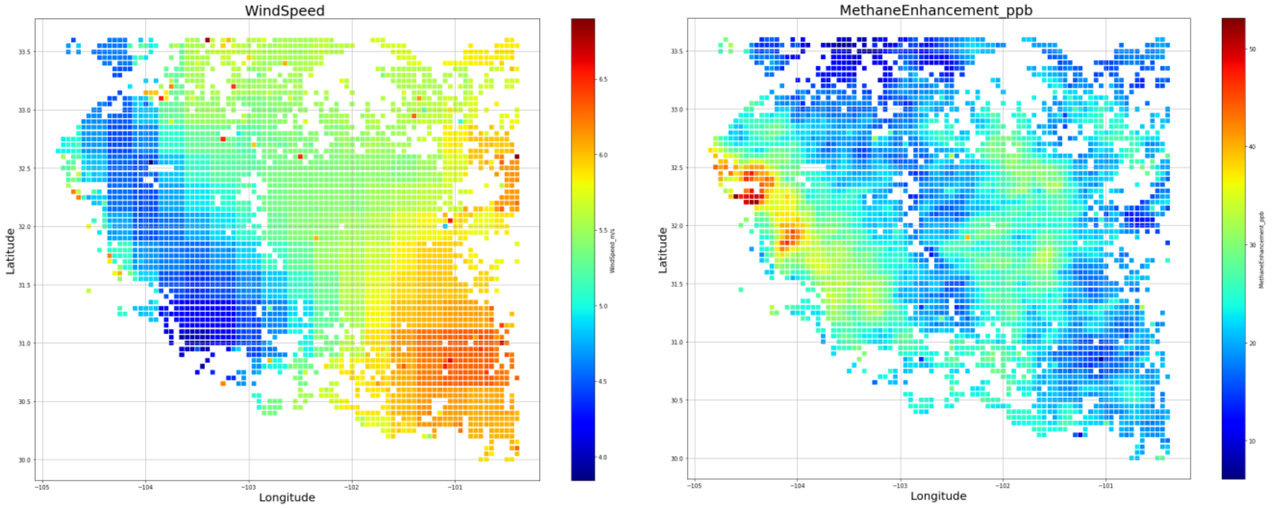
Figure 13. Feature importance ranking for the Random Forest Regression model using SHAP (SHapley Additive exPlanations) values and MDI (Mean Decrease Impurity).



(A)

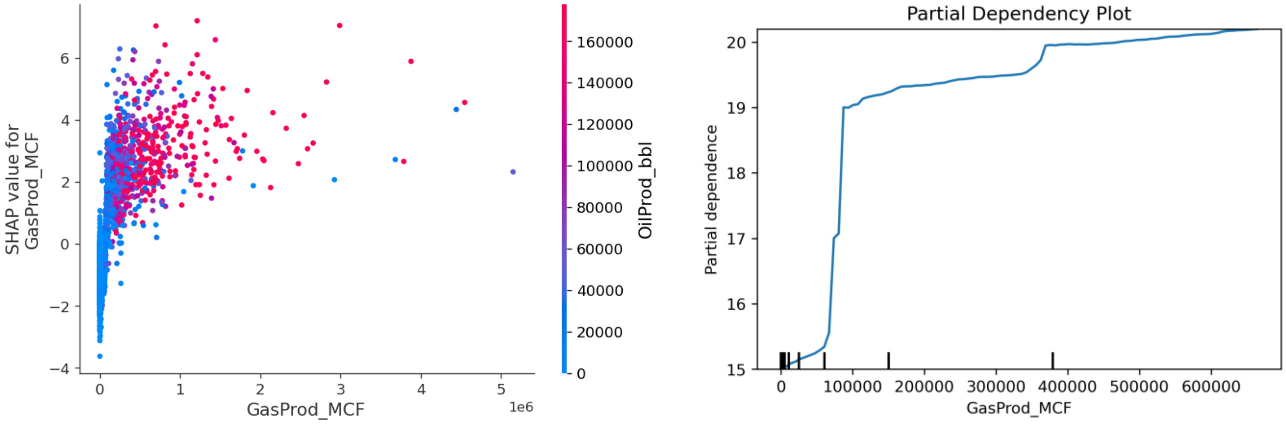
(B)

Figure 14. A) Monthly variation of windspeed in the Permian Basin. B) Monthly methane enhancement variation in the Permian Basin represented through the 25th, 50th, and 75th percentiles.



(A) (B)

Figure 15.) Spatial distribution of average wind speed over the Permian Basin. B) A) Spatial distribution of average methane enhancement over the Permian Basin.



(A) (B)

Figure 16. A) Contribution of Gas production to the Random Forest Regression model presented in SHAP independence plot. B) Partial Dependency Plot displays the effects of Gas production on the predicted methane enhancement on average.

Gas production emerges as an important variable influencing methane enhancements. Despite the Permian Basin's reputation as the most productive oil-producing region in the United States, the data suggests that gas production exhibits a more pronounced influence on methane emissions. An explicit positive correlation between gas production and methane enhancement is observed in Figure 16B. The partial dependency slope appears to be steepest in the range of approximately 0.8e5 MCF to 1e5 MCF of gas production, gradually tapering off at higher gas production levels. As the primary component of natural gas, methane can inadvertently be released into the atmosphere during various stages of production through several pathways, including, but not limited to, unintentional leakage, managed venting to balance system pressure, and flaring where methane is converted to CO_2, but some escapes due to insufficient conditions to sustain stable combustion (Johnson et al., 2011).

Wells with high gas production are also substantial oil producers, as evidenced by Figure 16A. However, wells that produce more oil may also possess a more comprehensive infrastructure to handle the produced gas than those lower-production wells. This may explain the observed reduction in the partial dependency slope at elevated levels of production seen in Figure 16B. Nevertheless, higher gas production increases the opportunity for methane emissions, leading to higher methane enhancement, as evidenced by the positive SHAP values. On the other hand, our analysis does not indicate a strong impact of oil production on methane enhancement (Figure 13).

Midstream facilities, including gas processing plants, gathering and boosting facilities, and transmission infrastructures, are potential sources of methane emissions. To evaluate the influence of facilities, particularly gas processing plants, on methane emissions, the proximity of every grid cell to the nearest such facility was assessed. Using gas processing plants as an example, we evaluated the influence of the distance to gas processing facilities utilizing both SHAP and Mean Decrease Impurity (MDI) metrics (Figure 17). Not surprisingly, the distance to the processing plant correlates inversely with methane enhancement. Additionally, it is observed that a lower surface altitude has a minor influence on methane enhancement, whereas regions of higher surface altitude register a more pronounced impact (Figure 17A). This impact can be either positive or negative. This observation also partially explains why surface altitude emerges as a more influential factor in the model than initially expected.

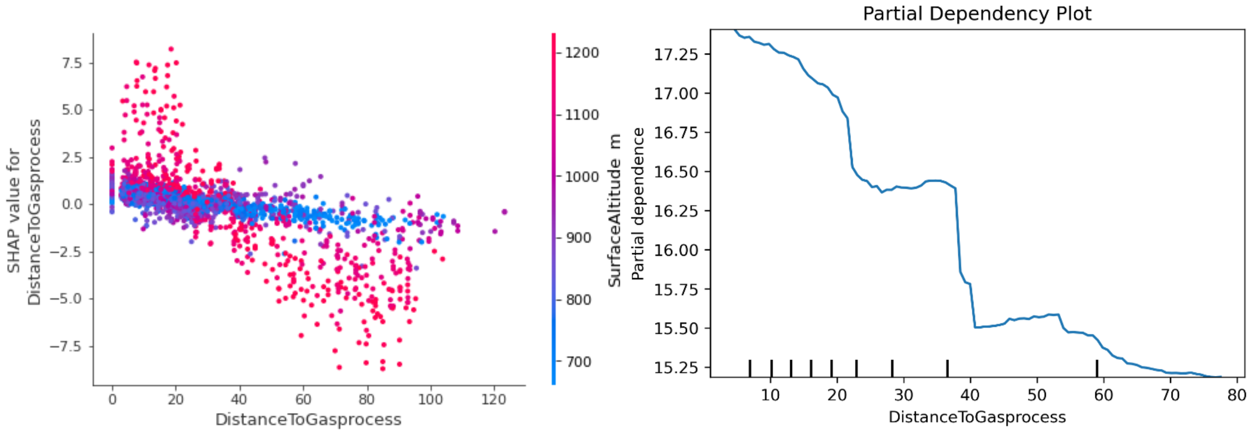


Figure 17. A) Contribution of Distance to Gas Processing facilities to the Random Forest Regression model presented in SHAP independence plot. B) Partial Dependency Plot displays the effects of Distance to Gas Processing facilities on the predicted methane enhancement on average.

Another influential feature in the model is landfill sites. Figure 18 illustrates the impact of distance to a landfill site on methane enhancement, as indicated by the SHAP values. The results indicate a notable positive impact within a 20-mile radius of landfill sites. However, beyond 50 miles, the impact diminishes and exhibits a negligible or negative effect on methane enhancement. The anaerobic decomposition of organic waste produces methane gas as a byproduct (IPCC, 2019). The methane gas generated in this process can escape from the landfill into the atmosphere, contributing to methane enhancements in the surrounding area. In some cases, landfill operators capture this gas and use it as a source of renewable energy or burn (flare) it off to reduce its greenhouse impact. A suspected reason behind its importance in the model is the continual emissions from landfills. Unlike certain oil and gas processes, which may operate cyclically or vary in intensity, landfills emit methane more regularly due to the constant decay of organic waste.

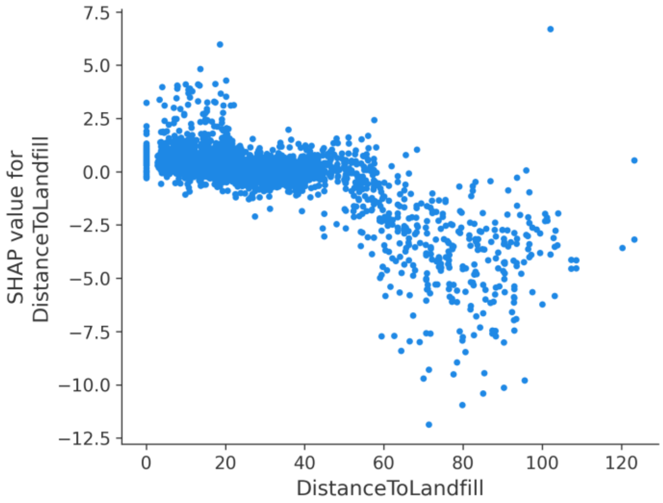


Figure 18. A) Contribution of Distance to Landfill sites to the Random Forest Regression model presented in SHAP independence plot.

4.3.3 Classification

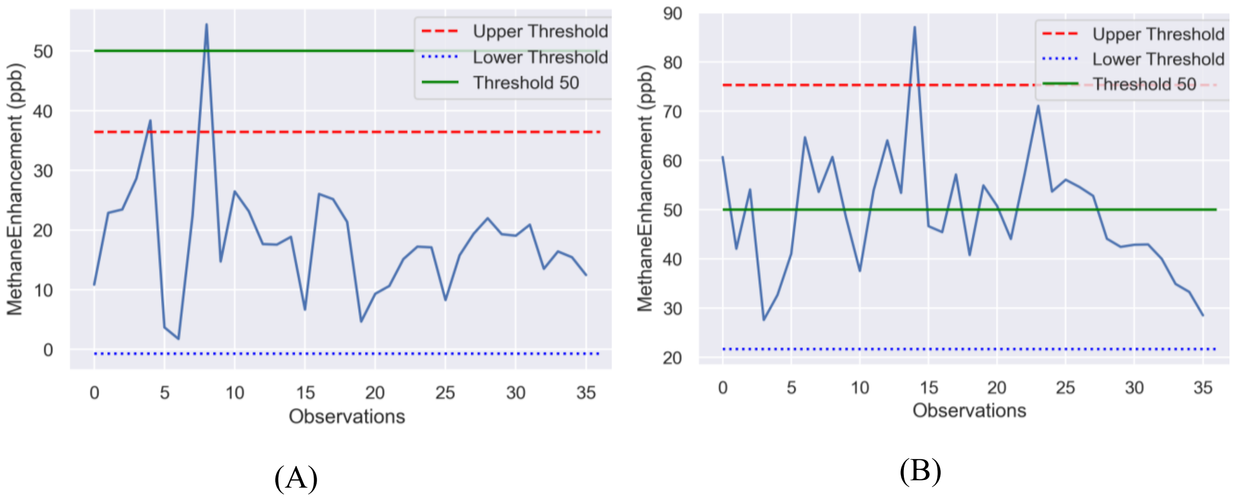
Predicting precise values of methane enhancement using a regression model presents significant challenges, mainly due to the low resolution of methane enhancement data sources and the unpredictability of fugitive emissions (outliers). Instead of predicting a precise methane enhancement, an alternative is to predict the occurrence of high, low, and normal emission levels.

4.3.3.1 Classification Threshold

Every methane enhancement data point at a grid cell is classified as high, normal (medium), or low according to the two criteria. First, considering all the historical measurements at that particular grid location, a data point exceeding the upper threshold – the third quartile plus 1.5 times the interquartile range (IQR) – is labelled as high, while a data point below the lower threshold – the first quartile minus 1.5 times the IQR – is labelled as 'Low,' aligning with the statistical definition of outliers. Second, all data points ≥ 50 ppb are labelled as 'High.'

A few remarks should be noted. A classification of 'High' does not necessarily represent an abnormal state of methane emissions. Rather, this classification is context-specific and reflects a high enhancement relative to this particular grid cell's historical data. On the other hand, a methane enhancement level deemed high for one grid cell might not be considered as such for another cell.

This approach ensures that the designation of high enhancement is contextually applicable and sensitive to the unique historical trend of each grid cell. For instance, as shown in Figure 19A, the respective grid cell demonstrates a historically lower level of methane enhancement. In this scenario, the upper threshold—calculated from the historical data of this specific grid cell—serves as the benchmark for distinguishing high enhancement. Conversely, the grid cell shown in Figure 19B exhibits a consistent trend of high methane enhancement, whereby the upper threshold surpasses the 50 ppb by a significant margin. The established 50 ppb threshold is employed for this grid cell to demarcate high methane enhancement.



(A)

(B)

Figure 19. Historical variation in methane enhancement for two distinct grid cells at A: (-101.4, 33,25) and B: (-104.4, 32.25), showing their upper and lower thresholds, respectively.

These classification labels or 'flags' serve a dual purpose: Firstly, any geographical locations exhibiting abnormal methane enhancement values are labelled as 'high' (Figure 20), facilitating easy visualization of methane emission status across various regions. Secondly, the flagged instances of anomalous methane enhancement provide essential guidance to the workings of the classification model. These classes allow for a deeper understanding of the various factors influencing methane emissions, enabling the model to distinguish between different types of enhancement scenarios.

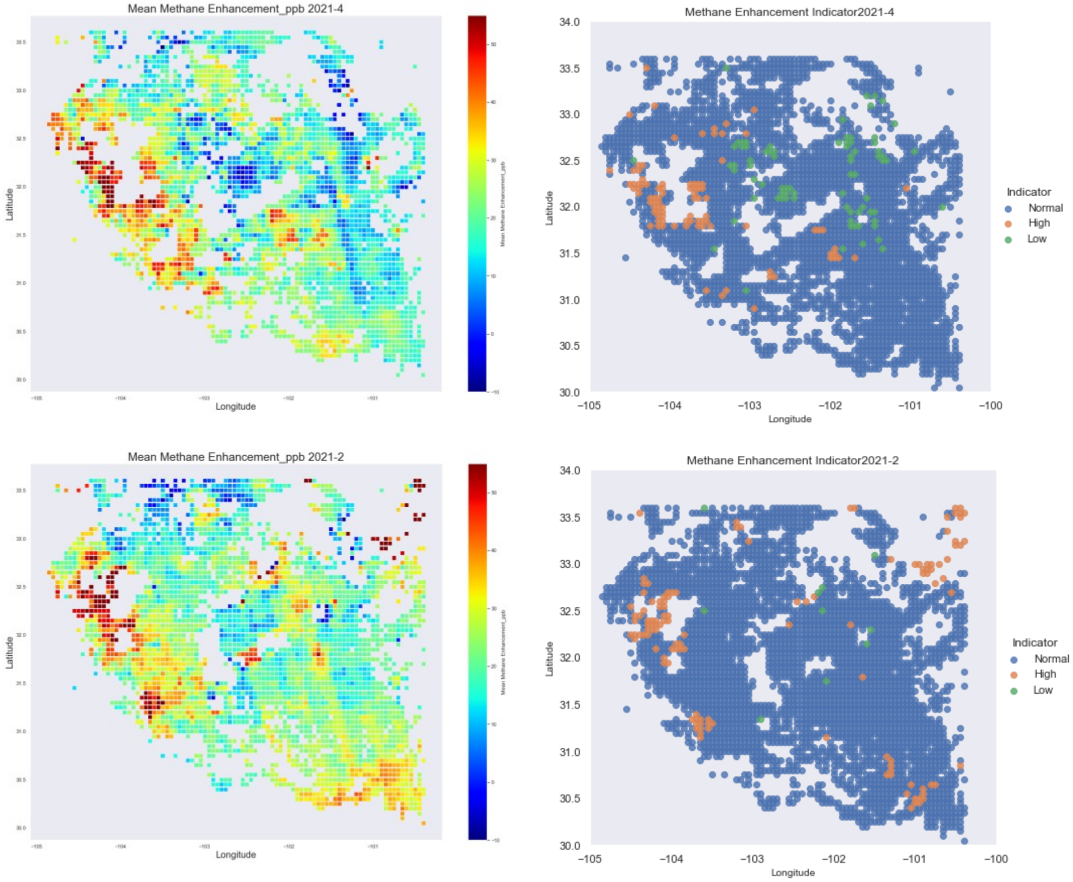

Mean Methane Enhancement_ppb 2021-4
Latitude
Longitude
Mean Methane Enhancement_ppb
Methane Enhancement Indicator2021-4
Latitude
Longitude
34.0
33.5
33.0
32.5
32.0
31.5
31.0
30.5
30.0
−105
−104
−103
−102
−101
−100
Indicator
Normal
High
Low
Mean Methane Enhancement_ppb 2021-2
Latitude
Longitude
Mean Methane Enhancement_ppb
Methane Enhancement Indicator2021-2
Latitude
Longitude
Indicator
Normal
High
Low

Figure 20. Monthly methane enhancement heat maps of the Permian Basin from 2021-2 and 2021-4 (Left) and their corresponding classifications (Right).

4.3.3.2 Random Forest Classification

The Random Forest classification model was selected, primarily due to its better performance relative to other models in the preceding regression analysis (Section 4.3.2). Given that the 'Normal' class cases significantly outnumber the 'High'"" and '"Low"' class cases, the model could be biased toward predicting the "Normal" class. Ensemble methods like Random Forest can more effectively handle imbalanced datasets. The prediction accuracy for the '"High",' '"Normal",' and '"Low"' classes (true positive rates) are 75%, 96%, and 20%, respectively, with a robust mean cross-validation score of 0.79.

Even though the model has a relatively modest true positive rate (75%) for the 'High' class prediction, many false positives are noted – instances from the 'High' class are predicted as the

'Normal' class. This factor contributes substantially to a reduced F1-score for the 'High' class, which currently stands at 0.14. The low F1-score indicates that the model may be unreliable in identifying areas with high methane enhancement. The sample size corresponding to the 'Low' class is considerably small, indicating the lesser emphasis on the prediction of low methane enhancement. Therefore, the low true positive rate associated with the prediction of the 'Low' class is acceptable, given that the primary focus of our research is not oriented towards accurately predicting instances of low emissions. The outcomes derived from the model performance metrics align with the findings observed in the previous regression model. Specifically, predictability tends to diminish at both extremes of methane enhancement – notably high or low levels – while it remains relatively robust for intermediate enhancement levels. This pattern highlights the challenges in modeling extreme methane enhancement levels under the current data resolution.

The Receiver Operating Characteristic (ROC) curve is another tool for assessing classification model performance. Each point plotted on the ROC curve corresponds to a unique pair of sensitivity and specificity values derived from a specific decision threshold adopted by the model (Fawcett, 2006). The curve demonstrates the trade-off between sensitivity (true positive rate) and specificity (one minus false positive rate) for every possible decision threshold (Fawcett, 2006). The Area Under the ROC curve (AUC) is an aggregate measure of the model's capacity to accurately distinguish among various classes (Bradley, 1997). An AUC value approaching unity indicates superior classification ability, which signifies that the model exhibits a high true positive rate and a low false positive rate across all thresholds. Therefore, a ROC curve hugging the upper left corner of the plot, yielding a larger AUC, reflects a model of higher overall accuracy.

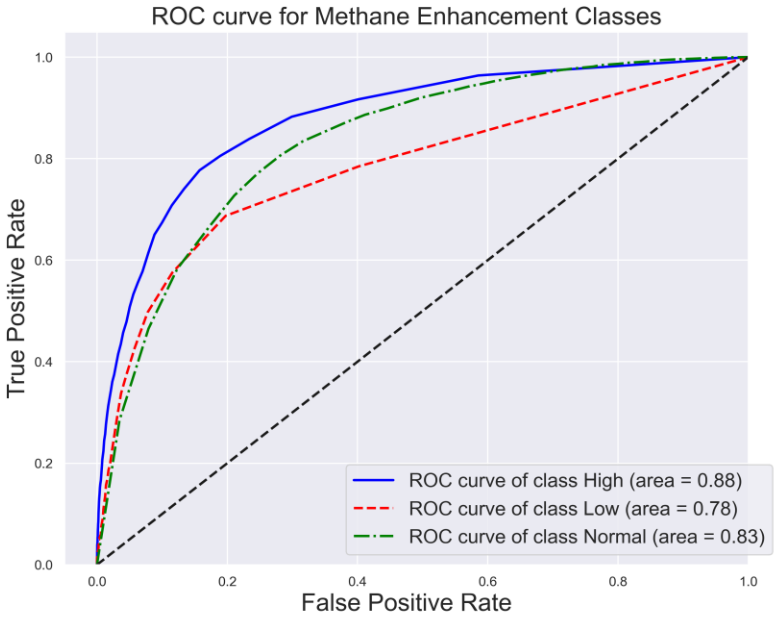


Figure 21. The Receiver Operating Characteristic (ROC) curves from the Random Forest Classification model.

Figure 21 illustrates that the model is better at predicting observations classified as 'High' compared to those classified as 'Normal' and 'Low.' The discrepancy between the AUC-ROC and other performance metrics (precision, recall, F1-score, etc.) may suggest that the model's performance is more complex than a single metric can capture. Different measures are sensitive to different aspects of the model's behavior. The AUC-ROC measures the overall ability of the model to discriminate between positive and negative classes (Saito, 2015). A high AUC indicates that the model has a good balance of sensitivity and specificity across different thresholds, which implies the model's good performance in differentiating the "High" class from the others. On the other hand, the F1-score is the harmonic mean of precision and recall (Sasaki, 2007). Precision is the proportion of true positive predictions out of all positive predictions, and recall (or sensitivity) is the proportion of true positive predictions out of all actual positives. The low F1-score and high AUC-ROC on the "High" and "Low" minority classes indicate that this model performs well at differentiating these instances but struggles with making the final classification. The AUC-ROC is not sensitive to imbalanced datasets as it considers the rank of predictions rather than their absolute values (He, 2009). However, the F1-score is heavily influenced by the model's performance on the minority "High" and "Low" classes, as it considers both precision and recall.

In summary, the observed discrepancies are largely due to imbalanced datasets, a well-recognized challenge in machine learning (He, 2009; Chawla, 2002; Krawczyk, 2016).

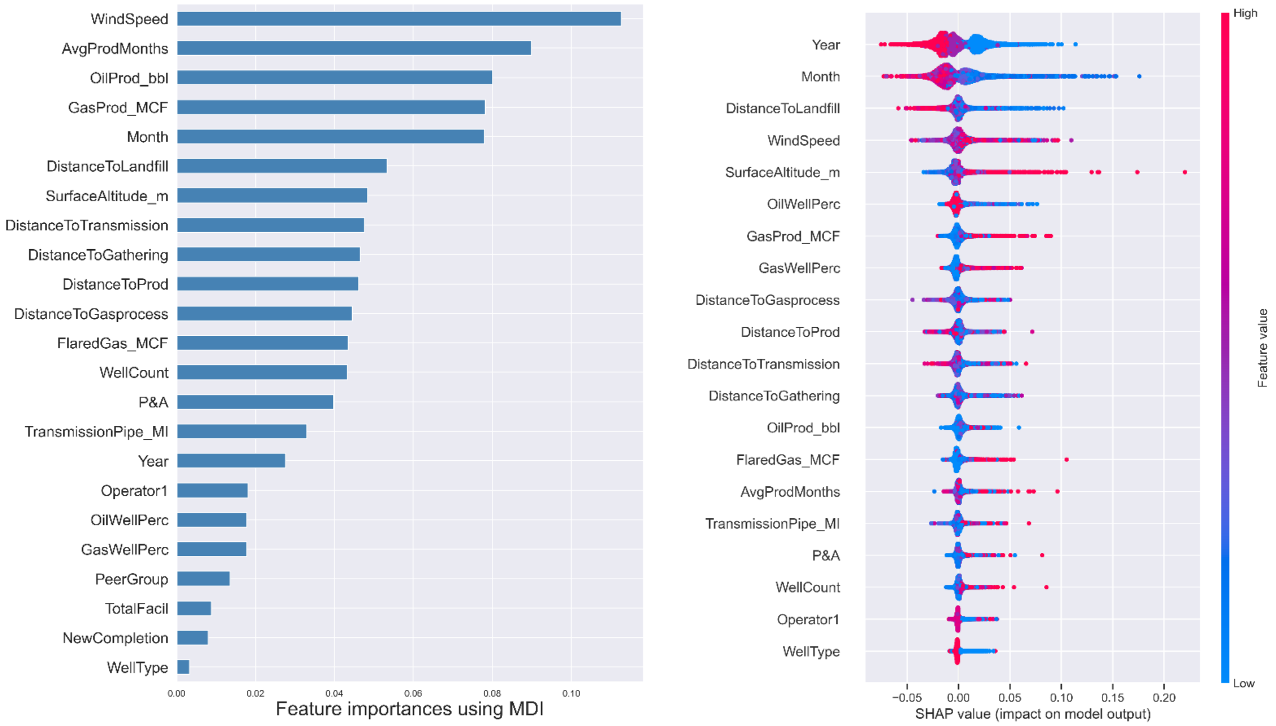
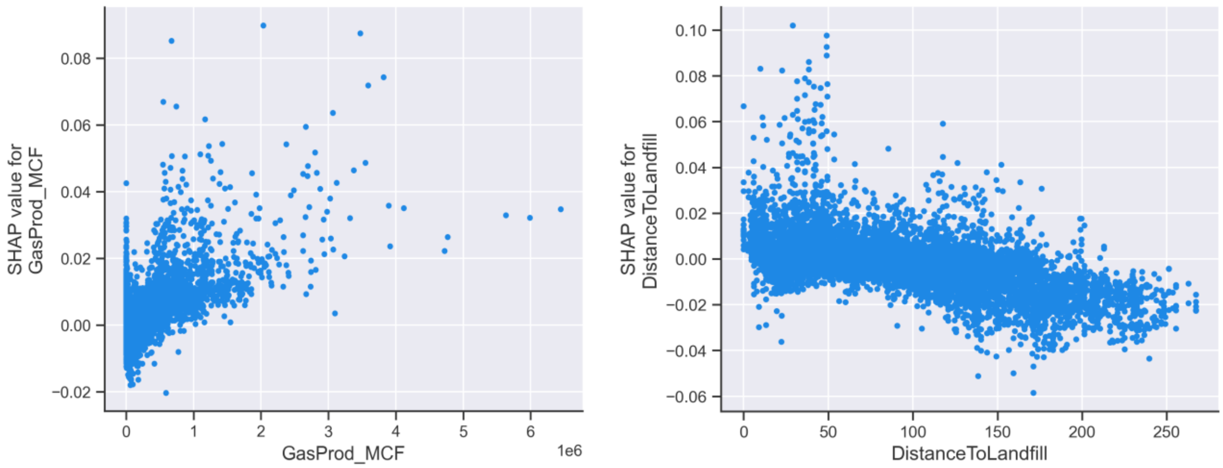


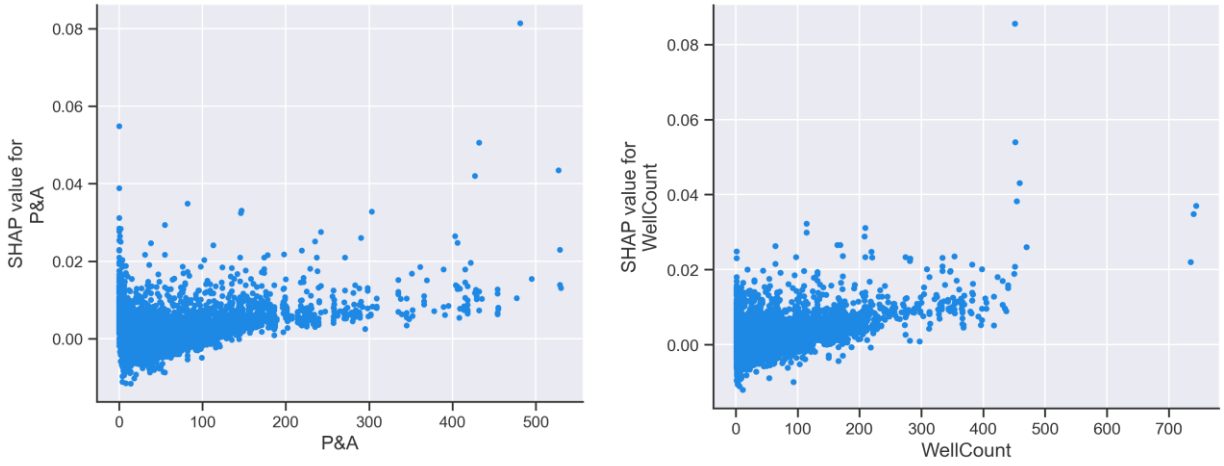
Figure 22. Feature importance ranking for the Random Forest Classification model in predicting 'High' class using MDI (Mean Decrease Impurity) and SHAP (SHapley Additive exPlanations) values.

Figure 22 shows that many of the most impactful features are similar to those in Figure 5 for the regression model. The SHAP values in the classification model are notably smaller than the regression model. However, despite the smaller SHAP values, the dependency plots of certain features exhibit clearer trends in the classification (Figure 23). The smaller SHAP values in the classification model indicate the feature's influence is relatively minor, implying that while the feature may not be a major driver in determining the "High" prediction on its own, it captures the distinct pattern or relationship that aligns with the predicting class. For example, the individual impact of the number of Plugged and Abandoned (P&A) wells within a grid cell, as well as the number of operating wells within a grid cell, are plotted in Figure 23C and 23D, respectively. A noticeable positive trend is observed in both plots, implying a proportional association between the possibility of abnormally high methane enhancement and more P&A and operating wells. This positive correlation was not noticeable in the regression model constructed using the entire dataset (Section 4.3.2), without special consideration of the variance in environmental and operational factors locally. However, the classification model compares the methane enhancement for each grid cell against its respective historical methane enhancement levels; this relative comparison within each grid cell offers a more standardized frame of reference, thereby increasing the ability to identify anomalies in methane enhancements. Considering each grid cell's historical context facilitates a form of normalization according to local baseline conditions and operational practices, enhancing the comparability and interpretability of observed methane enhancement classes. Thus, employing such a methodology increases the probability of identifying abnormal methane emission patterns.

The Permian Basin has had a long history of oil and gas development since the 1930s "("Permian Basin (North America")," Wikipedia). As a result, it is dotted with over 158 thousand P&A wells and over 46 thousand undocumented old wells (Enverus, PRSIM). Improperly decommissioned wells are susceptible to fugitive emissions. A higher count of these wells within a grid cell translates into an elevated risk of fugitive and operational emissions, increasing the possibility of abnormally high methane enhancement.



(A) (B)



(C)

(D)

Figure 23. Contribution of Gas Production Volume (A), Distance to Landfill sites (B), number of Plugged and Abandoned (P&A) wells (C) and the number of producing wells (D) in the Random Forest Classification model in predicting 'High' class presented in SHAP independence plot.

4.3.4 Clustering Analysis

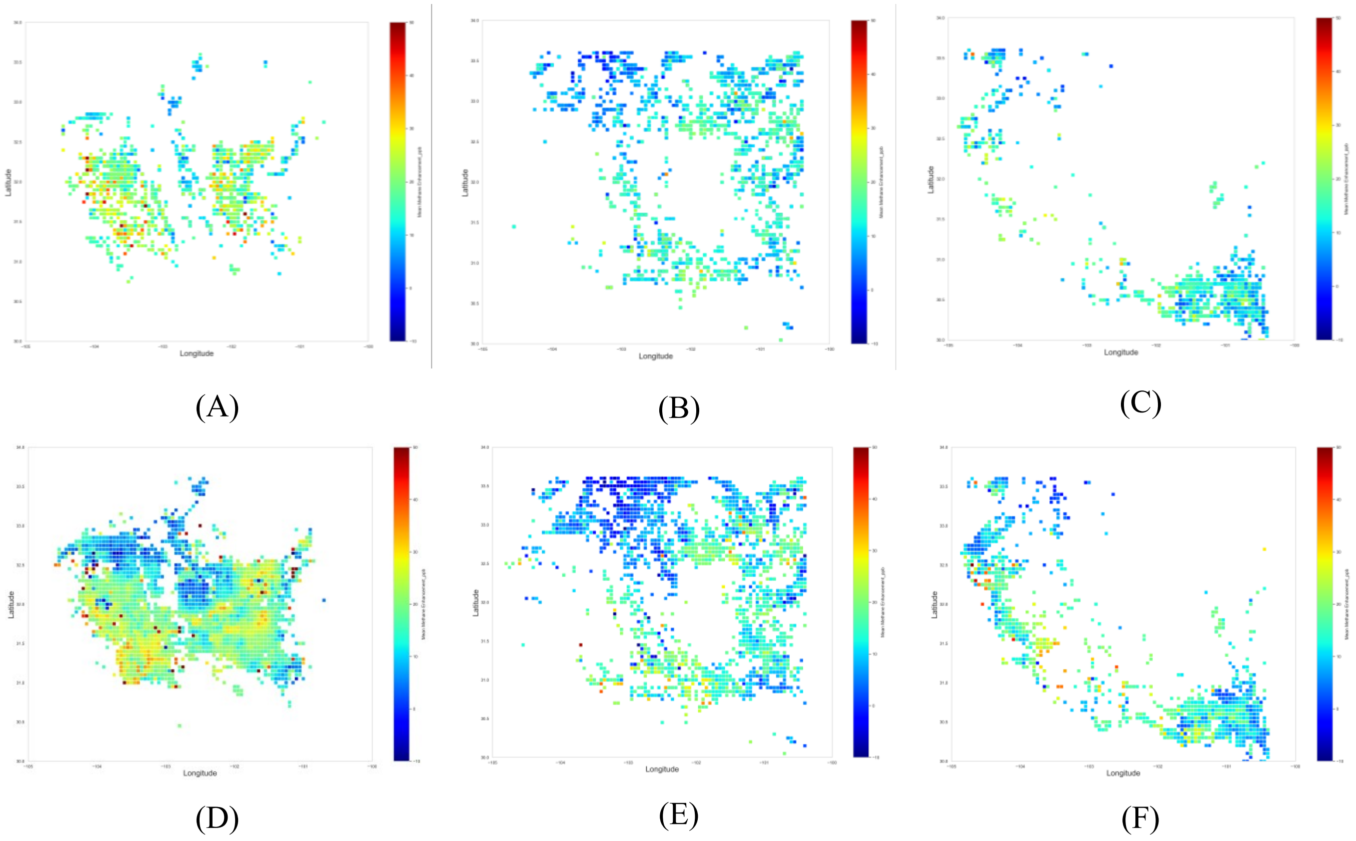
The previous analyses reveal the inherent complexity and the presence of internal structures in the dataset. This limitation comes from the often unpredictable nature of these emissions, limited data (e.g., operating conditions, regulatory practices, and maintenance procedures are not publicly available and absent from the data), and the difference in scale between individual well data, satellite measurements, and other environmental data, making the precise prediction and characterization of emission sources particularly challenging. In this section, clustering analysis is applied to segment the data into distinct groups, each characterized by a stronger internal homogeneity of features. This approach has the potential to provide a better understanding of what conditions are associated with certain methane emission behaviour.

Based on previous analysis and the characteristics of the dataset being used, the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm was selected for its inherent advantages in addressing this type of problem (Mclnnes et al., 2017). HDBSCAN is a density-based clustering algorithm with the unique feature of accommodating clusters of varying densities, distinguishing it from conventional distance-based clustering algorithms such as K-means (MacQueen, 1967). While K-means offers the benefits of speed and interpretability, HDBSCAN excels in identifying outliers by categorizing data points that are too sparse to belong to any specific cluster as noise. This feature significantly enhances its application in outlier detection. Further, its ability to discover clusters of varying densities makes HDBSCAN better at dealing with high-dimensional datasets.

To make the high-dimensional data more manageable, Principal Component Analysis (PCA) (Jolliffe, 2002) was employed, reducing the dimensionality from 31 to 20, yet preserving 95% of the data's variance. Following the application of HDBSCAN and careful parameter tuning, 53% of data points were classified as noise, with the remaining 47% comprising three distinct clusters. The high proportion of data classified as noise indicates that the patterns of methane emission are complex and cannot be easily grouped into clear clusters based on the existing variables at the current granularity. It could also imply that a more detailed dataset might be needed to capture the patterns of methane emissions. Higher granularity data might include, but not limited to, more specific measures of equipment types, more precise estimates of methane enhancements or more detailed geographical information. All these potential factors and their interactions could make isolating clear and discrete clusters within the data challenging.

Regarding the 47% of data points that were effectively clustered, it is noteworthy that despite excluding geographical coordinates as input for the clustering process, the resulting clusters seemed to reflect the geographical disposition of different areas within the Permian. Cluster 1 (Figure 24A) corresponded with the Delaware and Midland Basin, Cluster 2 (Figure 24B) represented the northern and eastern Permian and Central Basin Platform, and Cluster 3 (Figure 24C) was mainly aligned with the Val Verde Basin. This intriguing outcome potentially unveils spatially consistent patterns in methane enhancements across the Permian Basin.

For comparative purposes, the K-means++ algorithm (Arthur et al., 2007) was also applied to this dataset to partition it into three distinct clusters. The clusters formed (as shown in Figure 24 D-F) exhibit a remarkable similarity to those generated by the HDBSCAN algorithm, with each cluster coinciding with a distinct region within the Permian Basin. Unlike HDBSCAN, the K-means++ approach does not incorporate the identification of outliers or noise within the dataset. As a result, it assigns every data point to a particular cluster. This fundamental difference in K-means++ algorithm results in clusters that appear more comprehensive in their geographical representation than the ones from HDBSCAN. After looking into the hierarchical structure produced by HDBSCAN, it is observed that Cluster 1 (as displayed in Figure 24A) and Cluster 3 (Figure 24C) share many similarities and branch off from a common, larger cluster. Therefore, it was decided to merge these two clusters, thereby reducing the total number of distinct clusters to two. The revised clustering structure now represents a 'Main cluster', encompassing the Delaware, Midland, and Val Verde Basins, and a second cluster, referred to as the 'Other cluster'.

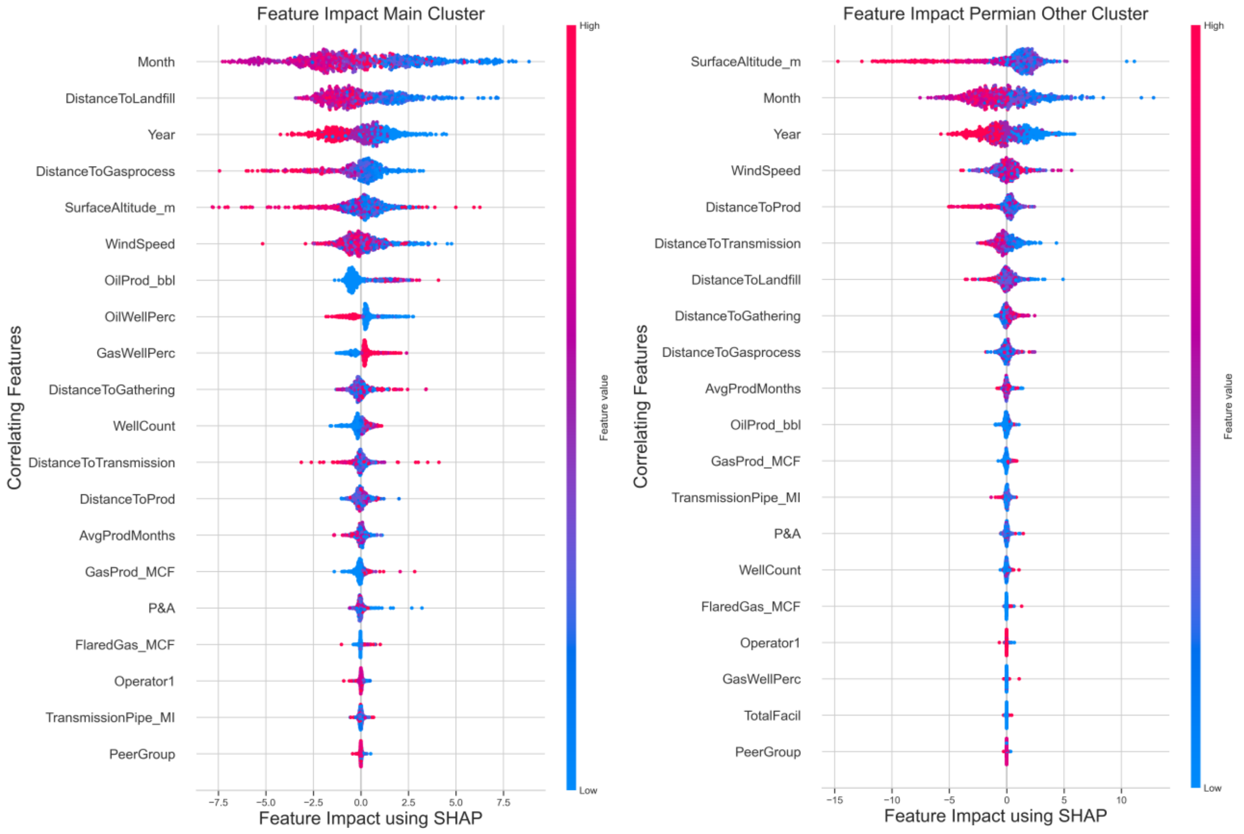


(A) (B) (C)

(D) (E) (F)

Figure 24. A-C): Clustering results from the HDBSCAN clustering algorithm. D-F): Clustering results from K-means++ clustering algorithm with 3 clusters.

We applied the Random Forest regression model independently to each of the two redefined clusters. Compared to the results in Section 4.3.2, a similar R² and lower RMSE (decreased by ~14% and declining to 5.70 ppb) are observed for the 'Main cluster.' The feature importance analysis within this Main cluster (Figure 25A) yields similar results as in Figure 5. Certain features, such as the Distance to Landfill Site (Figure 26A) and Distance to Gas Processing Facility (Figure 26B), have increased in significance within this cluster. The results here reinforce the conclusion that gas processing plants and landfill sites are significant contributors to methane emissions. Some other features, including Gas Well Percentage (Figure 26C) and Well Count (Figure 26D), demonstrate a more evident relationship regarding their impact on methane enhancement within this cluster. Grid cells with a greater percentage of gas wells and more operating wells tend to have a positive effect on methane enhancements. Additionally, we noticed that an increase in oil production, especially when accompanied by high gas production, has a greater impact on methane enhancement (Figure 26E). Wells producing high volumes of associated gas are often associated with higher methane emissions.

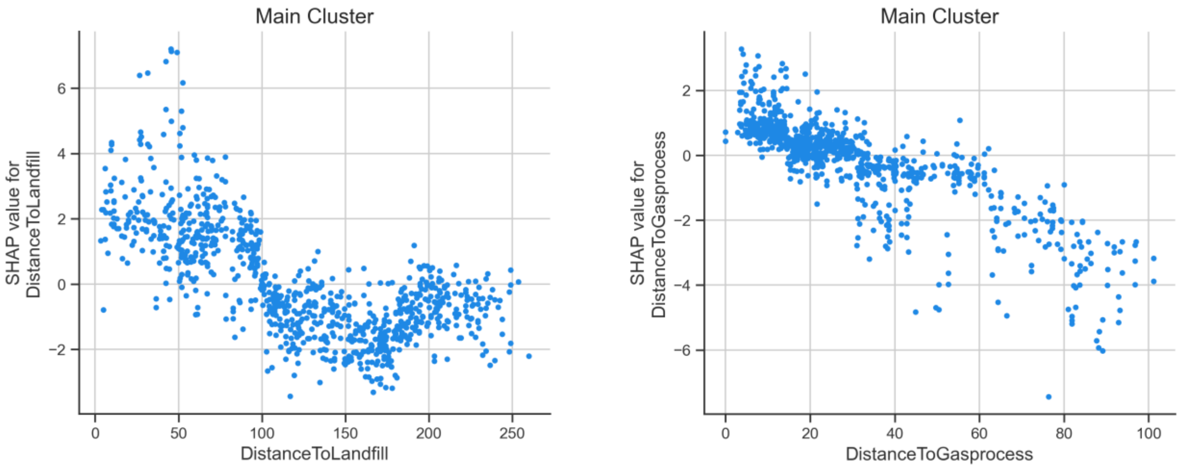


(A) (B)

Figure 25. Feature importance ranking for the Random Forest Regression model within the 'Main' cluster (A), and the 'Other' cluster (B) using SHAP (SHapley Additive exPlanations) values.

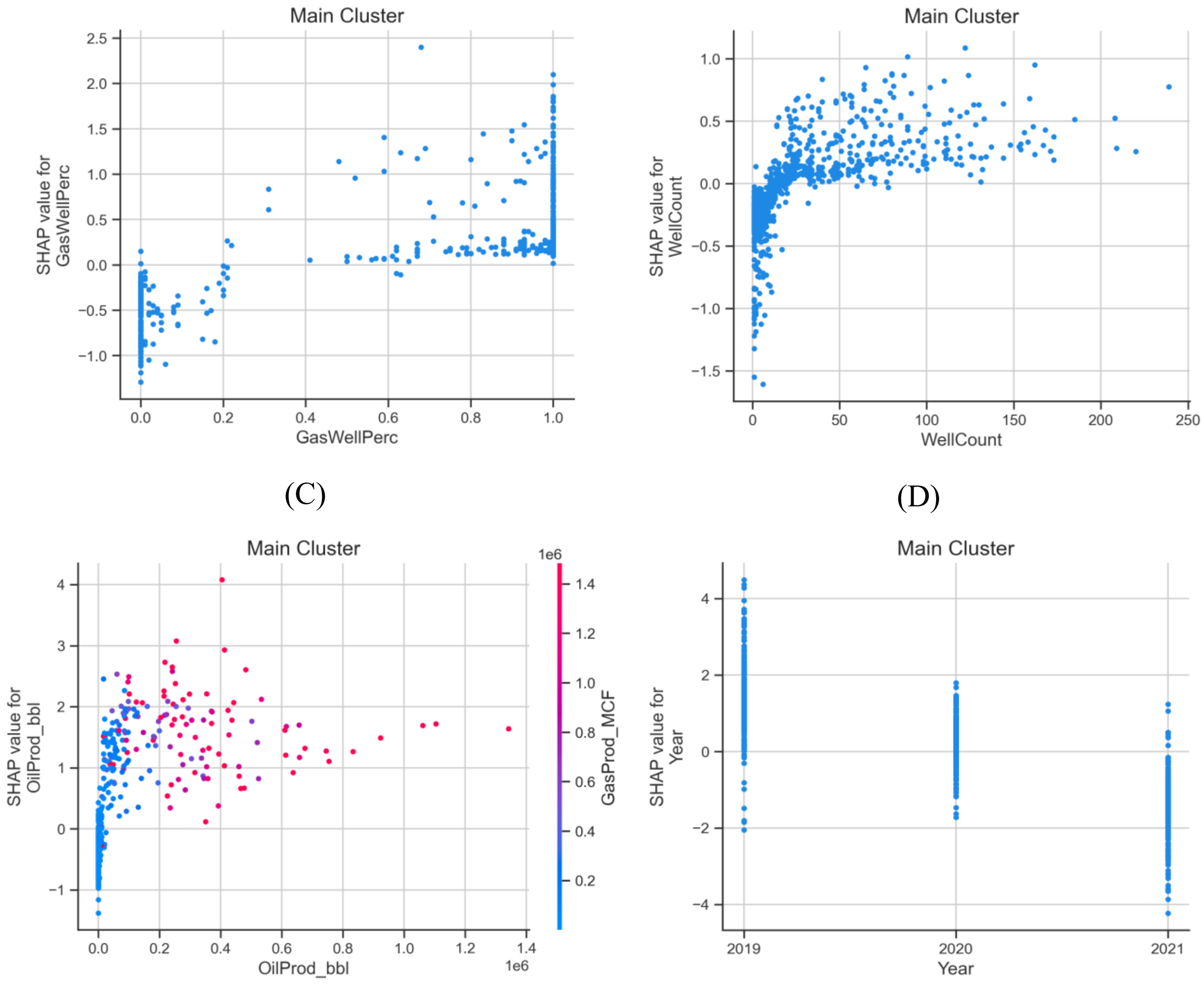
In the Random Forest regression model for the 'Other' cluster, the R² is calculated to be 0.56, accompanied by an RMSE of 6.38 ppb. The results from the feature importance analysis for this model are shown in Figure 25B. Surface Altitude is found to be the most influential feature and functioning as a proxy for geographic location, which exhibits a strong correlation with methane enhancement (as illustrated in Figure 27 A-B).

Interestingly, the significance of features relating to distance is more prominent in this cluster than in other cases we have analyzed. The feature 'Distance to Production' (Figure 27 C-D) is defined as the distance to any operator's nearest active production centroid. Similarly, 'Distance to Transmission' (Figure 27 E-F) refers to the distance to the closest compressor station. Both features symbolize the proximity to a location with active production, indicating potential exposure to emissions from these operations. This cluster principally covers geographic regions characterized by a low density of wells and reduced production activities. Therefore, it is reasonable to hypothesize that being closer to these active production centers increases its exposure to their emissions. In essence, the emissions from these centers have a broader-reaching impact due to the sparse nature of the regions within this cluster. Consequently, the measures of distance to these production centers become critical parameters in estimating methane enhancement.



(A)

(B)



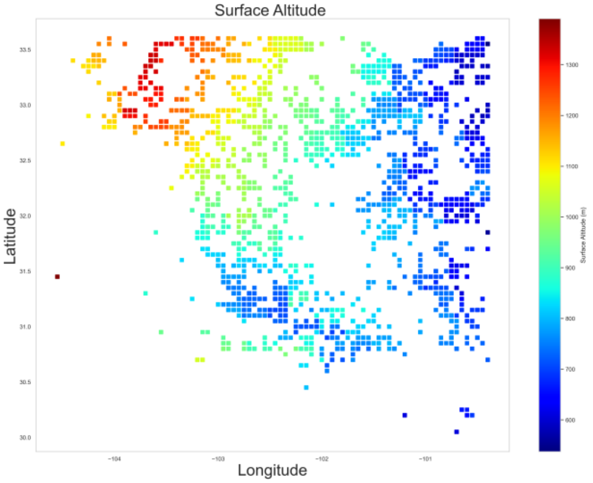
(C)

(D)

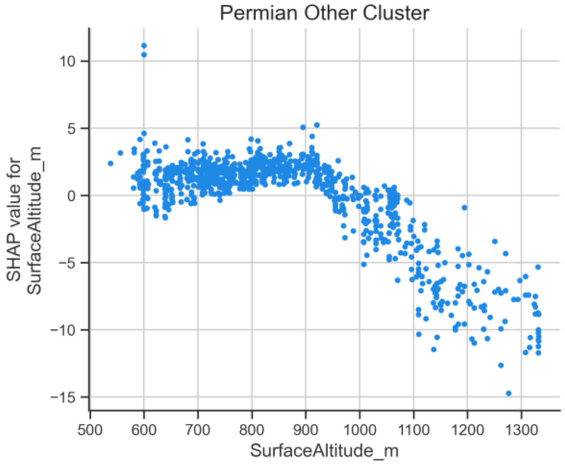
(E)

(F)

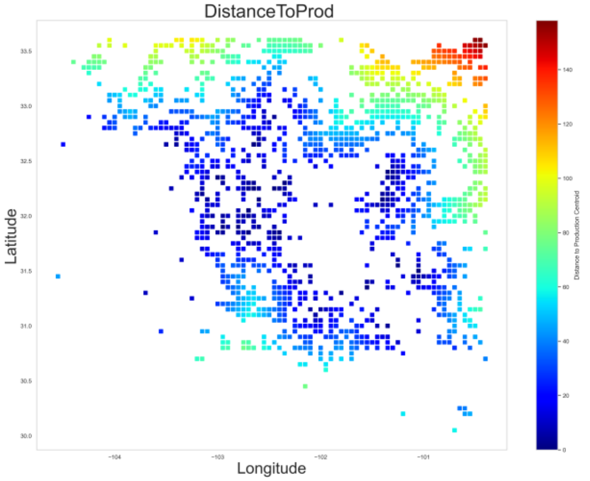
Figure 26. Contribution of Distance to Landfill sites (A), Distance to Gas Processing facilities (B), Gas well Percentage in the grid cell (C), the number of producing wells (D), Oil Production Volume (E), and the producing year in the Random Forest Regression model within the Main cluster presented in SHAP independence plot (F).


Surface Altitude
Latitude
33.5
33.0
32.5
32.0
31.5
31.0
30.5
30.0
Longitude
−104
−103
−102
−101
1300
1200
1100
1000
900
800
700
600
Surface Altitude (m)

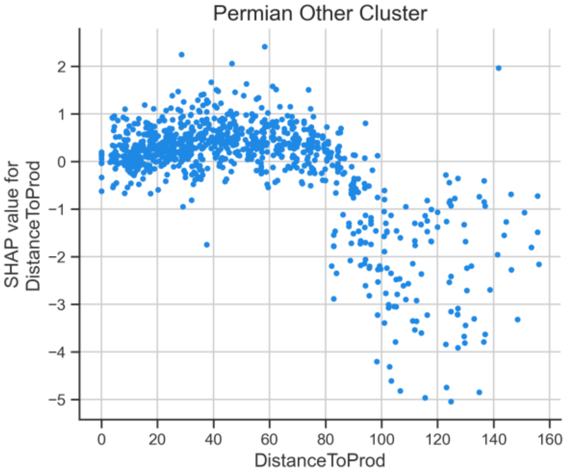
(A)


Permian Other Cluster
SHAP value for SurfaceAltitude_m
10
5
0
−5
−10
−15
500
600
700
800
900
1000
1100
1200
1300
SurfaceAltitude_m

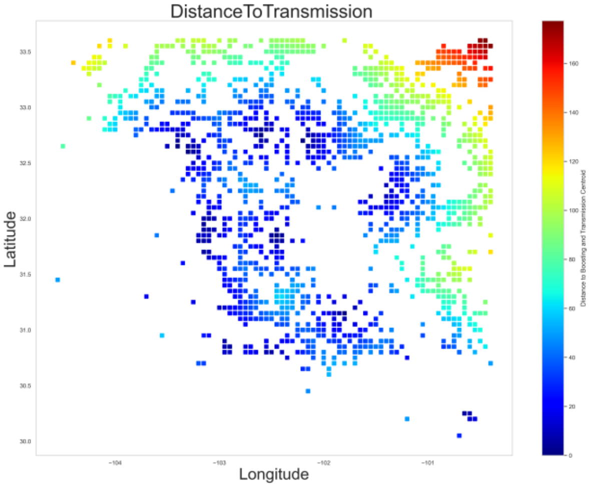
(B)


DistanceToProd
Latitude
33.5
33.0
32.5
32.0
31.5
31.0
30.5
30.0
Longitude
−104
−103
−102
−101
140
120
100
80
60
40
20
0
Distance to Production Centroid

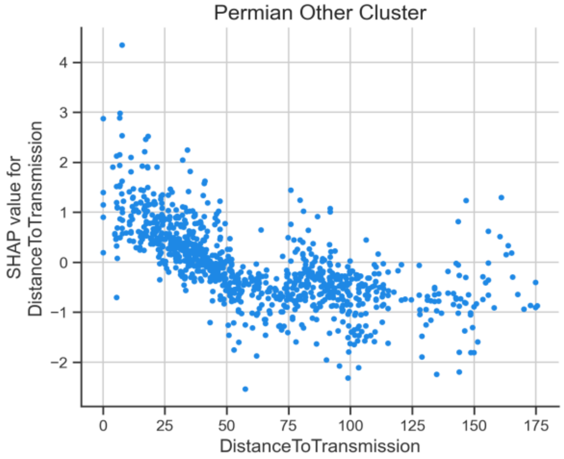
(C)


Permian Other Cluster
SHAP value for DistanceToProd
2
1
0
−1
−2
−3
−4
−5
0
20
40
60
80
100
120
140
160
DistanceToProd

(D)


DistanceToTransmission
Latitude
33.5
33.0
32.5
32.0
31.5
31.0
30.5
30.0
Longitude
−104
−103
−102
−101
160
140
120
100
80
60
40
20
0
Distance to Boosting and Transmission Centroid

(E)


Permian Other Cluster
SHAP value for DistanceToTransmission
4
3
2
1
0
−1
−2
0
25
50
75
100
125
150
175
DistanceToTransmission

(F)

Figure 27. Contribution of Surface altitude (A-B), Distance to nearest Production centroid (C-D), Distance to nearest compressor stations (E-F) in the Random Forest Regression model within the Permian Other cluster presented in SHAP independence plot (Right) and their corresponding geospatial heat maps (Left).

CHAPTER 5: CONCLUSIONS & FUTURE WORK

5.1 Summary and Conclusions

This research targeted two main objectives: Developing a practical workflow for regional methane monitoring and quantification, and understanding contributing factors to methane emissions in the oil and gas industry.

The workflow presented in Chapter 3 can facilitate practical regional monitoring and efficient estimation of methane emissions using daily data retrieved from the TROPOMI Sentinel-5P instrument. This workflow also integrates an interpolation approach to impute the missing values for better utilization of all high-quality data. A detailed analysis of two major hydrocarbon plays in the U.S. is presented as case studies to illustrate the workflow's feasibility and highlight the intrinsic characteristics, uncertainties, and application potential of the datasets.

The result shows that the estimated annual average rate of methane emission in the time period of 2019-2021 for the Appalachian Basin (4.92, 4.32, 4.14 Mt/year) surpasses the emission from the Permian Basin (4.46, 3.42, 2.84 Mt/year) in the corresponding years. This result is compared with the basin-wise annual rate of emission reported by a few other authors (Zhang et al., 2020; Sadavarte et al., 2021; Schneising et al., 2020). Possible factors contributing to the observed discrepancies may include variations in the temporal scope of the analysis, differing sizes of the monitored regions, and the potential for overestimation inherent in our methodology. The main sources of emissions in the Permian Basin are located in the Delaware and Midland Basins. In the Appalachian Basin, the retrieved data quality is imperfect due to its complex geographic settings resulting from less-than-ideal retrieval conditions. The emission rate from the Appalachian Basin is evaluated over two sub-regions (NE and SW). The average annual emissions from 2019 to 2021 show a decreasing trend for both sub-regions in the Appalachian Basin and the Permian Basin. This reduction indicates notable progress in the efforts made in recent years to mitigate methane emissions, marking a substantial improvement compared to the escalating trend observed between 2010 to 2015 in the United States (Sheng et al., 2018).

In chapter 4, the Permian data gathered from point emission sources are upscaled to the S5p resolution to explore complexity of methane emissions. Data analytics techniques are leveraged to correlate upscaled methane-related parameters with S5p methane measurements. While these emissions are difficult to model accurately due to their inherent variability and unpredictability, applying a series of analytical methods has provided significant insights into this complex issue.

The observed differences in emissions among different operator classes in the Permian Basin suggest influential factors beyond production volume. The differences in emissions amongst peer groups, especially between private and public companies and micro-cap operators, emphasize the potential role of operational practices, technologies used, and regulatory compliance in methane emissions. The regression and classification models identified key features that correlate with methane enhancements, including wind speed, month, surface altitude, gas production, and proximity to operational facilities and landfill sites. The findings reveal that a complex interplay of geographical, temporal, and operational parameters influences methane emissions. However, the study acknowledged the limitations of these models due to the multifaceted nature of methane emissions. Clustering analysis is employed to provide a richer understanding of the conditions under which methane emissions occur. Both the HDBSCAN and K-means++ algorithms are utilized for clustering analysis. Each provided valuable insights: The geographical disposition of the clusters identified by both algorithms, even without geographic coordinates as input, demonstrates the spatially consistent patterns in methane enhancements across the Permian Basin. The application of the Random Forest regression model on each of the redefined clusters revealed distinct factors influencing methane emissions within each cluster. These variables include oil production, proximity to landfill sites, gas processing plants, and well count in the Main cluster to proximity to active production centers in the 'Other cluster'.

5.2 Contributions

The research successfully designed and implemented an efficient workflow to monitor and estimate regional methane emissions using data from the TROPOMI Sentinel-5P instrument. The workflow integrates a novel interpolation approach to address missing data, improving the use of high-quality data for emissions estimation. This contribution advances current capabilities in methane monitoring and provides a practical solution for regional analysis.

This research inovates by applying data analytics techniques, such as regression and classification models and clustering analysis, to understand the complexity of methane emissions. This application led to the identification of key emission-enhancing features, demonstrated the potential role of operational practices, and revealed the spatial consistency of methane enhancements, thereby enriching the understanding of methane emissions.

This research underscores the importance of a tailored, context-specific approach to mitigating methane emissions. Given the diversity and heterogeneity of factors impacting these emissions, the analysis would likely be more effective if they were catered to the specifics of each region's emission patterns. While higher granularity data could enhance future analyses, this study provides a strong foundation for further research and intervention strategies. Through its analysis, this study contributes to a better understanding of methane emissions in the Permian Basin, serving as an important step toward developing more effective emission reduction strategies in the oil and gas industry. These findings will inspire further studies and more comprehensive data collection efforts, leading to a more refined understanding of relationships between operational features and methane emissions.

5.3 Future Work

There are several potential directions in terms of future work. One potential area could be refining the interpolation approach for imputing missing values in the data, to further improve the effectiveness of the workflow and the accuracy of methane emissions estimation. This could involve developing machine learning-based algorithms tailored for the specific characteristics of methane emissions data. Another one could involve further research into the impact of operational, geographic, and temporal parameters on methane emissions. As this research showed, these factors have a complex and significant impact on methane emissions. However, more work is needed to fully understand these relationships and their implications for emissions reduction strategies.

Expanding the geographic scope of the study is another potential avenue for future research. Applying the workflow developed in this study to different regions could reveal further insights into regional differences in methane emissions, leading to a more nuanced understanding of global methane dynamics.

The analysis presented in Chapter 4 demonstrates the utility of the S5p for tracking methane emissions on a global scale, although its limited resolution challenges the accurate identification of specific sources. Newer technologies on the horizon, such as hyperspectral remote sensing, hold promising potential to significantly augment the current capabilities of remote sensing technologies. This advanced technology is expected to offer a high-resolution view of the atmosphere and surface, substantially improving the precision of emission source detection and characterization. Furthermore, upcoming satellites such as MethaneSAT (Chan Miller et al., 2022), GOSAT-GW (NIES, 2021), Carbon Mapper (Duren et al., 2021), and CO2M (Sierk et al., 2021) are anticipated to provide significant advancements in methane emission remote sensing. These advancements include higher spatial and temporal resolution, increased accuracy, and wider spectral coverage, all of which outpace some of the current technologies. It is projected that the workflow developed in this research could be adapted to incorporate these novel technologies, providing a more powerful tool for greenhouse gas emissions monitoring and source characterization. While this research has made significant contributions in methane emissions monitoring and understanding, it also opens the door to many exciting opportunities for further exploration and innovation.

Bibliography

[1] Alvarez RA, Zavala-Araiza D, Lyon DR, Allen DT, Barkley ZR, Brandt AR, Davis KJ, Herndon SC, Jacob DJ, Karion A, Kort EA, Lamb BK, Lauvaux T, Maasakkers JD, Marchese AJ, Omara M, Pacala SW, Peischl J, Robinson AL, Shepson PB, Sweeney C, Townsend-Small A, Wofsy SC, Hamburg SP. Assessment of methane emissions from the U.S. oil and gas supply chain. Science. 2018 Jul 13;361(6398):186-188. doi: 10.1126/science.aar7204. Epub 2018 Jun 21. PMID: 29930092; PMCID: PMC6223263.

[2] Apituley, A., Pedergnana, Mattia et al.; Sentinel-5 Precursor/TROPOMI Level 2 Product User Manual Methane. Source: SRON; ref: SRON-S5P-LEV2-MA-001; issue: 0.11.6; date: 2019-06-24

[3] Arthur, D.; Vassilvitskii, S. (2007). "k-means++: the advantages of careful seeding", Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics Philadelphia, PA, USA. pp. 1027–1035.

[4] Bian, Jeffrey Y., Leung, Juliana Y., Volkmer, Nick, and Jingwen Zheng. "An Improved Workflow in Mass Balance Approach for Estimating Regional Methane Emission Rate Using Satellite Measurements." Paper presented at the SPE Canadian Energy Technology Conference and Exhibition, Calgary, Alberta, Canada, March 2023. doi: https://doi.org/10.2118/212791-MS

[5] BloombergNEF, The oil and gas industry's methane problem in four charts (2022, August 10). Retrieved January 10, 2023, from: https://about.bnef.com/blog/the-oil-and-gas-industrys-methane-problem-in-four-charts/

[6] Blöschl, G., & Sivapalan, M. (1995). Scale issues in hydrological modeling: A review. Hydrological Processes, 9(3-4), 251-290. doi: 10.1002/hyp.3360090305

[7] Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition, 30(7), 1145-1159. DOI: 10.1016/S0031-3203(96)00142-2

[8] Breiman, L. Random Forests. *Machine Learning* **45**, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

[9] Burrows, J. P., and Coauthors, 1999: The Global Ozone Monitoring Experiment (GOME): Mission Concept and First Scientific Results. *J. Atmos. Sci.*, **56**, 151–175, https://doi.org/10.1175/1520-0469(1999)056<0151:TGOMEG>2.0.CO;2.

[10] Butz, A., O. P. Hasekamp, C. Frankenberg, and I. Aben (2009), Retrievals of atmospheric CO2 from simulated space-borne measurements of backscattered near-infrared sunlight: Accounting for aerosol effects, Appl. Opt., 48, 3322–XXXX, doi:10.1364/AO.48.003322.

[11] Cambaliza, M. & Shepson, Paul et al. (2013). Assessment of uncertainties of an aircraft-based mass-balance approach for quantifying urban greenhouse gas emissions. Atmospheric Chemistry & Physics Discussions. 13. 29895-29945. 10.5194/acpd-13-29895-2013.

[12] Chan Miller, C., et al., Methane retrieval from MethaneAIR using the CO2 Proxy Approach: A demonstration for the future MethaneSAT mission, submitted to Atmos. Meas. Tech., 2022.

[13] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

[14] Coalbed Methane Outreach Program. https://www.epa.gov/cmop

[15] de Gouw, J.A., Veefkind, J.P., Roosenbrand, E. et al. Daily Satellite Observations of Methane from Oil and Gas Production Regions in the United States. Sci Rep 10, 1379 (2020). https://doi.org/10.1038/s41598-020-57678-4

[16] Duren, R.M., et al., Carbon Mapper: on-orbit performance predictions and airborne prototyping, AGU Fall Meeting, New Orleans, 2021.

[17] Dlugokencky, E. J., et al. (2003). "Observational constraints on recent increases in the atmospheric CH4 burden." Geophysical Research Letters, 30(19), 1984.

[18] D. Wunch, G. C Toon, J.-F. L Blavier et al.; The Total Carbon Column Observing Network. Philos. T. R.Soc. A.; 369 (2011) (1943), 2087; doi:10.1098/rsta.2010.0240.

[19] D. Wunch, G. C. Toon, P. O. Wennberg et al.; Calibration of the Total Carbon Column Observing Network using aircraft profile data. Atmospheric Measurement Techniques; (2010), 1351; doi:10.5194/amt-3-1351-2010. DOI: 10.1021/acs.estlett.2c00380

[20] Dubey, L.; Cooper, J.; Hawkes, A. Minimum detection limits of the TROPOMI satellite sensor across North America and their implications for measuring oil and gas methane

emissions, Science of The Total Environment(2023). ISSN 0048-9697, https://doi.org/10.1016/j.scitotenv.2023.162222.

[21] ENVERUS. Prism. https://www.enverus.com/solutions/ energy-analytics/ep/prism/

[22] EPA (2019) Inventory of U.S. Greenhouse Gas Emissions and Sinks: 1990-2021. U.S. Environmental Protection Agency, https://www.epa.gov/ghgemissions/inventory-us-greenhouse-gas-emissions-andsinks-1990-2021.

[23] EPA (2023) Inventory of U.S. Greenhouse Gas Emissions and Sinks: 1990-2021. U.S. Environmental Protection Agency, https://www.epa.gov/ghgemissions/inventory-us-greenhouse-gas-emissions-andsinks-1990-2021.

[24] Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861-874. DOI: 10.1016/j.patrec.2005.10.010

[25] Frankenberg C, Meirink JF, van Weele M, Platt U, Wagner T. Assessing methane emissions from global space-borne observations. Science. 2005 May 13;308(5724):1010-4. doi: 10.1126/science.1106644. Epub 2005 Mar 17. PMID: 15774724.

[26] GHGSat. (2021). Global emissions monitoring. https://www.ghgsat.com/

[27] HARP 1.17, S[&]T, The Netherlands. http://stcorp.github.io/harp/doc/html/index.html

[28] Hasekamp, O., Lorente, Alba., Haili Hu et al.; Algorithm Theoretical Baseline Document for Sentinel-5 Precursor methane retrieval. Source: SRON; ref: SRON-S5P-LEV2-RP-001; issue: 1.10; date: 2019-02-01

[29] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263-1284. DOI:10.1109/TKDE.2008.239

[30] Hu, H.; Hasekamp, O.; Butz, A.; Galli, A.; Landgraf, J.; Aan de Brugh, J.; Borsdorff, T.; Scheepmaker, R.; Aben, I. The operational methane retrieval algorithm for TROPOMI. Atmos. Meas. Tech. 2016, 9, 5423-5440.

[31] International Energy Agency (2022), Methane Tracker Database, IEA, Paris.

[32] International Energy Agency. (2017). Energy access outlook 2017: From poverty to prosperity. Paris: OECD/IEA.

[33] IPCC. (2019). IPCC Guidelines for National Greenhouse Gas Inventories, Volume 5: Waste. Retrieved from https://www.ipcc-nggip.iges.or.jp/public/2019rf/index.html

[34] Javadinejad, S., Eslamian, S. & Ostad-Ali-Askari, K. Investigation of monthly and seasonal changes of methane gas with respect to climate change using satellite data. Appl Water Sci 9, 180 (2019). https://doi.org/10.1007/s13201-019-1067-9

[35] Jevan Yu, Benjamin Hmiel, David R. Lyon, Jack Warren, Daniel H. Cusworth, Riley M. Duren, Yuanlei Chen, Erin C. Murphy, and Adam R. Brandt. Environmental Science & Technology Letters 2022 9 (11), 969-974

[36] Johnson, M. R., & Coderre, A. R. (2011). An analysis of flaring and venting activity in the Alberta upstream oil and gas industry. Journal of the Air & Waste Management Association, 61(2), 190-200.

[37] Jolliffe, I. T. (2002). Principal Component Analysis (2nd edition). Springer.

[38] Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence, 5(4), 221-232.

[39] Kumar, P., Broquet, G., Caldow, C., Laurent, O., Gichuki, S., Cropley, F., ... & Ciais, P. (2022). Near-field atmospheric inversions for the localization and quantification of controlled methane releases using stationary and mobile measurements. Quarterly Journal of the Royal Meteorological Society.

[40] Kuze, A., Suto, H., Shiomi, K., Kawakami, S., Tanaka, M., Ueda, Y., Deguchi, A., Yoshida, J., Yamamoto, Y., Kataoka, F., Taylor, T. E., and Buijs, H. L.: Update on GOSAT TANSO-FTS performance, operations, and data products after more than 6 years in space, Atmos. Meas. Tech., 9, 2445–2461, https://doi.org/10.5194/amt-9-2445-2016, 2016.

[41] Landgraf, J., Lorente, A., et al. (2021). S5P MPC Product Readme Methane V02.03.01. Source: SRON; ref: S5P-MPC-SRON-PRF-CH4. issue 2.1; date: 2021-11-17

[42] Liu, M., van der A, R., van Weele, M., Eskes, H., Lu, X., Veefkind, P., et al. (2021). A new divergence method to quantify methane emissions using observations of Sentinel-5P TROPOMI. Geophysical Research Letters, 48, e2021GL094151. https://doi. org/10.1029/2021GL094151

[43] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In NIPS 2017: Proceedings of the 31st Conference on Neural Information Processing Systems (pp. 4765-4774). arXiv: 1705.07874

[44] Lyon, D. R., Hmiel, B., Gautam, R., Omara, M., Roberts, K. A., Barkley, Z. R., Davis, K. J., Miles, N. L., Monteiro, V. C., Richardson, S. J., Conley, S., Smith, M. L., Jacob, D. J.,

Shen, L., Varon, D. J., Deng, A., Rudelis, X., Sharma, N., Story, K. T., Brandt, A. R., Kang, M., Kort, E. A., Marchese, A. J., and Hamburg, S. P.: Concurrent variation in oil and gas methane emissions and oil price during the COVID-19 pandemic, Atmos. Chem. Phys., 21, 6605–6626, https://doi.org/10.5194/acp-21-6605-2021, 2021.

[45] MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. Vol. 1. University of California Press. pp. 281–297. MR 0214227. Zbl 0214.46201.

[46] Methane Guiding Principles. (n.d.). Reducing Methane Emissions Across the Natural Gas Value Chain. *http://www.methaneguidingprinciples.org/*

[47] McInnes, L, Healy, J, Astels, S (2017) hdbscan: Hierarchical density based clustering, Journal of Open Source Software 2(11), 205. https://doi.org/10.21105/joss.00205

[48] Muñoz Sabater, J., (2019): ERA5-Land hourly data from 1981 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). DOI: 10.24381/cds.e2161bac

[49] NIES, GOSAT-GW mission: background, aims, and mission requirements, 2021. https://gosat-gw.global-atmos-chem-lab.jp/en/collaboration/

[50] Pandey S, Gautam R, Houweling S, van der Gon HD, Sadavarte P, Borsdorff T, Hasekamp O, Landgraf J, Tol P, van Kempen T, Hoogeveen R, van Hees R, Hamburg SP, Maasakkers JD, Aben I. Satellite observations reveal extreme methane leakage from a natural gas well blowout. Proc Natl Acad Sci U S A. 2019 Dec 26;116(52):26376-26381. doi: 10.1073/pnas.1908712116. Epub 2019 Dec 16. PMID: 31843920; PMCID: PMC6936547.

[51] Pankaj Sadavarte, Sudhanshu Pandey, Joannes D. Maasakkers, Alba Lorente, Tobias Borsdorff, Hugo Denier van der Gon, Sander Houweling, and Ilse Aben. Methane Emissions from Superemitting Coal Mines in Australia Quantified Using TROPOMI Satellite Observations Environmental Science & Technology 2021 55 (24), 16573-16580. DOI: 10.1021/acs.est.1c03976

[52] Papritz, A.; Stein, A. (2002). "Spatial prediction by linear kriging". Spatial Statistics for Remote Sensing. Remote Sensing and Digital Image Processing. Vol. 1. p. 83. doi:10.1007/0-306-47647-9_6

[53] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

[54] Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLoS ONE 10(3): e0118432. DOI: 10.1371/journal.pone.0118432

[55] Sasaki, Y. (2007). The truth of the F-measure. Teach Tutor Mater., 1(5), 1-5.

[56] Schneising, O., Buchwitz, M., Reuter, M., Vanselow, S., Bovensmann, H., and Burrows, J. P.: Remote sensing of methane leakage from natural gas and petroleum systems revisited, Atmos. Chem. Phys., 20, 9169–9182, https://doi.org/10.5194/acp-20-9169-2020, 2020.

[57] Sentinel Hub. (n.d.). Sentinel-5P L2 Data - Sentinel Hub API documentation. Retrieved 2023-08-03, from https://docs.sentinel-hub.com/api/latest/data/sentinel-5p-l2/

[58] Shen, L., Gautam, R., Omara, M., Zavala-Araiza, D., Maasakkers, J., Scarpelli, T., Lorente, A., Lyon, D., Sheng, J., Varon, D., Nesser, H., Qu, Z., Lu, X., Sulprizio, M., Hamburg, S., Jacob, D., 2022. Satellite quantification of oil and natural gas methane emissions in the US and Canada including contributions from individual basins. Atmos. Chem. Phys. Discuss. 2022, 1–22

[59] Sheng, J.-X., Jacob, D. J., Turner, A. J., Maasakkers, J. D., Benmergui, J., Bloom, A. A., Arndt, C., Gautam, R., Zavala-Araiza, D., Boesch, H., and Parker, R. J.: 2010–2016 methane trends over Canada, the United States, and Mexico observed by the GOSAT satellite: contributions from different source sectors, Atmos. Chem. Phys., 18, 12257–12267, https://doi.org/10.5194/acp-18-12257-2018, 2018.

[60] Shepard, Donald (1968). "A two-dimensional interpolation function for irregularly-spaced data". Proceedings of the 1968 ACM National Conference. pp. 517–524. doi:10.1145/800186.810616

[61] Sierk, B., V. Fernandez, J.-L. Bézy, Y. Meijer, Y. Durand, G. Bazalgette Courrèges-Lacoste, C. Pachot, A. Löscher, H. Nett, K. Minoglou, L.Boucher, R. Windpassinger, A. Pasquet, D. Serre, and F. te Hennepe, The Copernicus CO2M mission for monitoring anthropogenic carbon dioxide emissions from space," Proc. SPIE 11852, International Conference on Space Optics — ICSO 2020, 118523M, doi: 10.1117/12.2599613, 2021.

[62] Tuzson, B., Graf, M., Ravelid, J., Scheidegger, P., Kupferschmid, A., Looser, H., ... & Emmenegger, L. (2020). A compact QCL spectrometer for mobile, high-precision methane sensing aboard drones. Atmospheric Measurement Techniques, 13(9), 4715-4731.

[63] United Nations. (n.d.). Net zero coalition. United Nations. Retrieved November 1, 2022, from https://www.un.org/en/climatechange/net-zero-coalition

[64] Varon, D. J.; Jacob, D. J.; McKeever, J.; Jervis, D.; Durak, B. O. A.; Xia, Y.; Huang, Y. Quantifying methane point sources from fine-scale satellite observations of atmospheric methane plumes. Atmos. Meas. Tech. 11, 5673–5686 (2018)

[65] Varon, D. J., McKeever, J., Jervis, D., Maasakkers, J. D., Pandey, S., Houweling, S., et al. (2019). Satellite discovery of anomalously large methane point sources from oil/gas production. Geophysical Research Letters, 46, 13,507–13,516. https://doi.org/10.1029/2019GL083798

[66] World Bank. (n.d.). Global Gas Flaring Reduction Partnership (GGFR). *https://www.worldbank.org/en/programs/gasflaringreduction*

[67] Yakovlev, S., Sadovnikov, S., & Romanovskii, O. (2022). Mobile Airborne Lidar for Remote Methane Monitoring: Design, Simulation of Atmospheric Measurements and First Flight Tests. Remote Sensing, 14(24), 6355.

[68] Y. Zhang, R. Gautam, S. Pandey, M. Omara, J. D. Maasakkers, P. Sadavarte, D. Lyon, H. Nesser, M. P. Sulprizio, D. J. Varon, R. Zhang, S. Houweling, D. Zavala-Araiza, R. A. Alvarez, A. Lorente, S. P. Hamburg, I. Aben, D. J. Jacob, Quantifying methane emissions from the largest oil-producing basin in the United States from space. Sci. Adv. 6, eaaz5120 (2020).

[69] "Permian Basin (North America)." Wikipedia, The Free Encyclopedia. Wikimedia Foundation, Inc. 10 April 2023. Permian Basin (North America) - Wikipedia

Appendices

Feature	**Explanation**
WindSpeed	Wind speed at the surface level
SurfaceAltitude_m	Altitude of the surface in meters
Month	Month of the year
OilProd_bbl	Oil production in barrels
GasProd_MCF	Gas production in thousand cubic feet (MCF)
AvgProdMonths	Average production duration in months
NewCompletion	Number of newly completed wells
Year	Year of observation
WellCount	Total number of wells
GasWellPerc	Percentage of gas wells
OilWellPerc	Percentage of oil wells
Operator1	Indicator variable for a specific operator
PeerGroup	Categorical variable representing the peer group of the operator
P&A	Number of plugged and abandoned wells
TotalFacil	Total number of facilities
TransmissionPipe_MI	Length of transmission pipelines in miles
DistanceToProd	Average distance to production facilities
DistanceToGasprocess	Average distance to gas processing facilities
DistanceToGathering	Average distance to gathering facilities
DistanceToLandfill	Average distance to landfill sites
DistanceToTransmission	Average distance to transmission facilities
FlaredGas_MCF	Volume of flared gas in thousand cubic feet (MCF)