MSc Business Analytics Consultancy Project/Dissertation 2024-25

CANDIDATE #	**MPZH0**
DATE	**11/08/2025**
WORD COUNT	**11334**
TITLE OF PROJECT	
Predictive Modelling for Methane (CH_4) Emissions Estimation	

Confidentiality Statement ☒

Where a project partner designates that a project should remain confidential (i.e. only made available to the supervisor & second marker), please check the box above. By doing so, this submission will remain as a record of academic work completed and will not be copied, reproduced, transferred, distributed, leased, licensed or shared with any other individual(s) and/or organisations, including web-based organisations other than the designated markers at any point in time.

Student Disclaimer:

I hereby declare that this dissertation is my individual work and to the best of my knowledge and confidence, it has not already been accepted in substance for the award of any other degree and is not concurrently submitted in candidature for any degree. It is the end product of my own independent study except where other acknowledgement has been stated in the text.

Criteria/Weight	Supervisor's comments
Topic, theoretical framework, literature & methodology (30%): • Topic is clearly identified, with well-defined boundaries. • Demonstrates knowledge of relevant theories and their limitations. • Uses current and relevant literature from reliable sources. • Develops an appropriate and adequate methodology for the topic. • Ensures methodology facilitates replication and reproducibility of results.	
Analysis and conclusions & recommendations (30%): • Uses primary and/or secondary data effectively. • Conducts rigorous analysis and interpretations. • Considers alternative interpretations/arguments. • Identifies and justifies limitations with reasonable arguments. • Draws conclusions/recommendations that are fully consistent with the evidence presented. • Demonstrates an understanding of the business context.	
GenAI Use and Critical Evaluation (10%): • Explores GenAI's potential for the project. • Tests and compares multiple GenAI tools/methods. • Evaluates AI-generated code, insights, literature reviews ...etc. • Documents when, where, and why GenAI is used. • Addresses ethical concerns such as data privacy, AI dependency, and academic integrity	
Structure, originality & presentation (10%): • Provides a concise and coherent summary. • Maintains a well-structured and logical presentation. • Demonstrates proper language use, style, graphs, tables, and referencing. • Uses appropriate and effective visualisations. • Presents meaningful business recommendations.	
Complexity of project scope & progress made towards business goals (10%): • Demonstrates progress in overcoming technical and operational challenges encountered during the project. • Shows advancement in addressing problem framing and data-related challenges.	
Project Management (10%): • Demonstrates structured project planning and management • Effectively engages with the supervisor throughout the dissertation process.	

General marking guidelines			
85 +	Outstanding work of publishable standard.	**50 - 59**	Good work which only covers a basic analysis. Some problems but no major omissions.
70 - 84	Excellent work showing mastery of the subject matter & excellent analytical skills.	**40 - 49**	Inadequate work. Not sufficiently analytical. Some major omissions.
60 - 69	Very good work. Interesting analysis with original insights. Some minor errors.	**0 - 39**	Work seriously flawed. Lack of clarity & argumentation. Too descriptive.

FINAL MARK: _____

Abstract

This dissertation develops a machine learning framework for predicting facility-level methane (CH_4) emissions across the United Kingdom, addressing the critical need for enhanced monitoring of short-lived climate pollutants. Conducted in collaboration with GreentecAI, a London-based environmental analytics company, the study integrates multi-dimensional datasets to create a scalable emissions forecasting system that supports both regulatory compliance and strategic mitigation planning. Point-source emissions data from the UK National Atmospheric Emissions Inventory (2018–2022) were combined with high-resolution gridded climate variables and spatial characteristics, producing a dataset of 18,130 facility-year observations across industrial sectors. The methodology involved spatial data fusion, seasonal climate aggregation, and strategic feature engineering to capture the complex interactions between environmental conditions, industrial activity, and emission patterns.

Machine learning algorithms such as Random Forest, and XGBoost were evaluated under different feature configurations. The optimal model, an Optuna-tuned XGBoost using a strategically selected subset of 14 features, achieved robust predictive performance with an R^2 of 0.72 and Mean Absolute Error of 0.27 on the holdout test set. SHAP (SHapley Additive exPlanations) analysis revealed that sectoral classification, spatial clustering metrics, and seasonal climate variables (notably winter temperature and wind speed) were the strongest predictors of methane emissions. Methane emissions displayed pronounced spatial and sectoral variability, with certain regions showing the highest per-facility emission intensity despite lower facility density. Four sectors were responsible for over 70% of total emissions, and facility clustering within 25 km correlated strongly and negatively with individual emission levels. This research demonstrates the feasibility of integrating heterogeneous spatial–temporal datasets for granular emissions forecasting. Practical recommendations include implementing SHAP-enhanced monitoring dashboards, prioritising inspections based on risk profiles, and developing sector-specific mitigation strategies. While limitations exist regarding temporal resolution and agricultural emissions coverage, the framework provides a robust foundation for operational deployment and extension to other greenhouse gases and regions.

Table of Contents

1. Introduction

1.1 Climate Context and CH_4

Amid growing global urgency to curb climate change, attention is increasingly shifting toward climate pollutants that can deliver rapid cooling benefits. One of the most powerful among these is methane (CH_4), a short-lived but highly potent greenhouse gas with over 80 times the warming power of carbon dioxide over a 20-year period (Mathew et al., 2024). This elevated short-term global warming potential means that, tonne-for-tonne, methane traps significantly more heat in the atmosphere than CO_2 in the decades immediately following its release. Because CH_4 also breaks down much faster than CO_2, persisting for just over a decade, cutting methane emissions can yield rapid climate benefits, slowing the rate of warming within a single human generation and buying time for longer-term decarbonisation efforts to take effect (Smeeton, 2022).

Methane is widely regarded as the second most significant greenhouse gas driving global warming, following carbon dioxide (Mathew et al., 2024). In the United Kingdom, major sources of CH_4 include agriculture, waste management, and the energy sector. These emissions are often highly variable, spatially dispersed, and harder to track in real time compared to carbon dioxide, presenting a significant challenge for policymakers and environmental agencies seeking to reduce national emissions in line with the UK's legally binding net-zero commitments (Smeeton, 2022).

Beyond national targets, methane reduction is increasingly viewed as a global opportunity to provide immediate mitigation potential in the race against irreversible climate tipping points. As a result, institutions such as the United Nations Environment Programme (UNEP), the Intergovernmental Panel on Climate Change (IPCC), and the European Environment Agency have emphasised the need for better CH_4 data resolution, forecasting, and intervention tools. Academic research has similarly begun to focus on enhancing methane monitoring systems using satellite data, ground-based sensors, and predictive models. However, much of this

work remains constrained by methodological gaps, particularly in integrating point-source emissions with climate and sectoral trends.

Accurate and timely emissions forecasting, especially for CH_4, is essential for driving effective mitigation strategies. This includes not only identifying and targeting high-emitting sites but also forecasting future risk patterns based on environmental conditions and operational activity.

1.2 GreentecAI and Project Context

This dissertation is developed in collaboration with GreentecAI, a London-based company that builds AI-powered digital ecosystems to accelerate sustainable transformation across infrastructure, energy, and environmental systems. The company delivers intelligent solutions combining software, data, and systems innovation for environmental analytics, energy management, and smart infrastructure.

GreentecAI's innovation lab is currently exploring how machine learning can enhance greenhouse gas monitoring capabilities, with a particular focus on methane due to its outsized short-term impact on global warming. As part of this broader effort, the company has developed a prototype model that estimates CH_4 emissions. However, this prototype faces scalability challenges, particularly when processing large climate datasets

This dissertation contributes to GreentecAI's goal of building a more robust and operationally viable solution by designing, implementing, and testing an improved forecasting pipeline. By integrating emissions data from multiple years, sector-wide national greenhouse gas summaries, and high-resolution gridded climate variables, the project will help the company develop a scalable model architecture that supports forecasting and decision-making.

1.3 Aim and Objectives

This dissertation aims to develop a robust CH_4 forecasting system for GreentecAI by advancing its current predictive modelling approach. While the company has already built a prototype emissions model, it is constrained by computational limitations.

The project addresses this challenge by exploring patterns in methane emissions across key UK sectors and time periods, transforming and aligning historical emissions, sectoral, and environmental data, and engineering features capable of supporting scalable machine learning models. These models will be trained and evaluated to forecast facility-level CH_4 emissions, and the results will be analysed to identify key drivers of methane variability.

The resulting system is intended to deliver a more robust and scalable forecasting solution that can operate effectively under real-world data constraints. The project will also explore options for creating a simple, user-friendly interface to test model predictions and allow for real-time data input, while ensuring that the final solution is designed to integrate with external systems to support future scalability and deployment. In doing so, the dissertation not only contributes to GreentecAI's immediate operational goals but also establishes a methodological foundation for emissions forecasting that can be expanded across other gases, regions, or use cases.

1.4 Report Structure

The remainder of this dissertation is structured into six chapters.

Chapter 2 presents a review of the academic and industry literature on methane emissions, spatial data integration, and the use of machine learning in environmental monitoring. It synthesises recent research on CH_4 forecasting, environmental sensor data, and emissions modelling frameworks, and identifies key methodological gaps that this project addresses.

Chapter 3 outlines the methodology used to acquire, clean, and combine emissions, sectoral, and climate datasets. It describes the feature engineering process, spatial matching techniques, and modelling pipeline including the rationale for model selection and evaluation metrics. This chapter also explains how the computational and integration challenges were addressed in the development phase.

Chapter 4 presents the results of the predictive models, evaluates performance metrics, and analyses feature importance and spatial emission patterns.

Visualisations, residual analysis, and error decomposition are used to assess the model's predictive power and practical usability.

Chapter 5 discusses the findings in the context of GreentecAI's objectives, highlights key limitations of the approach, and considers the broader implications for emissions forecasting, regulatory compliance, and sustainability strategy.

Chapter 6 concludes the dissertation and offers recommendations for model deployment, operational improvements, and future research directions, particularly in integrating additional data sources or expanding the modelling framework beyond CH_4.

Chapter 7 provides a critical reflection on the use of generative AI (GenAI) tools throughout the research process, including their role in literature sourcing, code development, and dissertation structuring.

2. Literature Review

2.1 Environmental Context and Theory

Methane (CH_4) is a potent greenhouse gas that plays a critical role in near-term climate change. Although its atmospheric concentration is far lower than that of carbon dioxide (CO_2), methane's global warming potential (GWP) is approximately 28–34 times higher than CO_2 over a 100-year period, making it a key target for climate mitigation strategies (IPCC, 2021). Furthermore, methane's relatively short atmospheric lifetime of about 9–12 years means that reductions in CH_4 emissions can result in near-immediate climate benefits, making it a highly effective lever for slowing the rate of global warming in the next few decades (Saunois et al., 2020). These attributes have led policymakers and researchers alike to prioritize methane reductions as an essential component of efforts to achieve the goals outlined in the Paris Agreement and other international climate frameworks.

In urban environments, methane emissions present unique challenges due to the diversity and complexity of emission sources. Point sources, such as landfills, wastewater treatment plants, and natural gas distribution networks, contribute a substantial share of city-wide CH_4 emissions (Vollrath et al., 2024). These sources are often highly concentrated geographically but can vary in intensity over time due to operational practices, infrastructure maintenance schedules, and technological differences across cities. In addition, diffuse emissions from transportation systems, household energy use, and other distributed sources, while individually small, can cumulatively contribute to urban CH_4 concentrations and are often more difficult to monitor and quantify accurately. This variability in both space and time highlights the inherent complexity of methane dynamics in densely populated regions and underscores the need for advanced monitoring and predictive tools.

Environmental factors further complicate efforts to monitor and understand urban methane emissions. Meteorological conditions, including wind speed, atmospheric pressure, and temperature inversions, play a critical role in determining how methane disperses and accumulates within the urban boundary layer (Stecher et al., 2025). For instance, low wind speeds and temperature inversions can trap methane near the surface, leading to transient concentration spikes or "hotspots" that may not be captured by standard monitoring techniques. Seasonal variations in temperature

and precipitation patterns further exacerbate these dynamics. Warmer summer temperatures, for example, have been linked to increased biogenic methane production in landfills and wetlands due to enhanced microbial activity, intensifying urban methane emissions during certain times of the year (Chang et al., 2025). Such dependencies on meteorological and climatic variables make it clear that any effective emissions monitoring strategy must incorporate environmental drivers into its design.

Traditional methane monitoring systems, however, often fall short in capturing these complexities. Bottom-up inventories, which are constructed using estimated emission factors and activity data, frequently lack the spatial and temporal resolution necessary to detect short-lived emission events or to account for small, distributed sources (Saunois et al., 2020). These limitations can lead to significant underreporting of emissions, particularly in older urban infrastructures where fugitive leaks from natural gas pipelines may occur unpredictably. On the other hand, top-down approaches, including satellite-based remote sensing platforms have dramatically improved global methane detection capabilities but remain constrained by coarse spatial resolution, limited revisit frequencies, and susceptibility to cloud cover interference (Vollrath et al., 2024). While such methods are invaluable for global and regional assessments, they are often inadequate for capturing the fine-grained variability present within cities.

This tension between the strengths and weaknesses of existing monitoring methods has led to increasing interest in hybridized approaches that combine ground-based measurements, atmospheric modelling, and advanced data analytics to create more robust urban methane monitoring systems. Integrating diverse data sources in this way enables a more nuanced understanding of emissions patterns and supports the development of targeted mitigation strategies. By emphasizing the interplay between urban infrastructures, environmental drivers, and technological constraints, the literature highlights a critical need for innovation in monitoring frameworks that can adapt to the dynamic nature of urban methane emissions and provide actionable insights for policymakers and urban planners.

2.2 Machine Learning in Environmental Science

The application of machine learning in environmental science has grown substantially in recent years, offering innovative approaches for greenhouse gas monitoring and prediction. Traditional emissions monitoring systems often rely on static models and manual reporting, which can be resource intensive and lack the responsiveness required for real time decision making. In contrast, machine learning techniques enable the analysis of large and complex datasets and support adaptive prediction systems that reflect the dynamic nature of environmental processes.

A systematic review highlights the increasing use of tree-based models such as Random Forest and gradient boosting machines for estimating carbon emissions (Alnuaimi et al., 2025). These models are particularly effective in handling high dimensional data and nonlinear relationships, which are common in environmental systems. While tree-based models often achieve strong predictive performance, careful feature selection and validation are essential to avoid overfitting, especially when applied to heterogeneous datasets such as point source emissions and climate grids. The review also emphasizes the advantage of tree-based methods in providing interpretable outputs that are valuable for policymakers and environmental managers.

In the agricultural context, Random Forest models have been successfully applied to predict methane and nitrous oxide emissions from diversified cropping systems (Bista et al., 2025). These models captured complex interactions between soil properties, climate variables, and management practices, achieving R squared values exceeding 0.85. The study also demonstrated the importance of incorporating domain knowledge during model development to select relevant features and reduce model complexity. However, the authors observed that the predictive accuracy of Random Forest models is highly dependent on the quality and resolution of the input data, reinforcing the need for robust preprocessing and rigorous validation in environmental machine learning applications.

Deep learning approaches have also gained traction in recent studies. One investigation combined convolutional neural networks and long short-term memory architectures to create a hybrid model for near real time greenhouse gas monitoring (Hasan et al., 2025). This system, which fused satellite imagery with ground-based

Internet of Things sensor networks, achieved a detection accuracy of 95 percent and reduced reporting latency from 24 hours to one hour. These results demonstrate the potential of deep learning for integrating spatial and temporal data streams in emissions prediction. Nevertheless, challenges such as the computational intensity of training and the limited interpretability of deep learning models remain critical considerations for operational deployment.

A long-term evaluation of machine learning based predictive emissions monitoring systems compared six algorithms including adaptive boosting and artificial neural networks for nitrogen oxide emission prediction (Si et al., 2024). Moderate complexity models like adaptive boosting achieved a favourable balance between robustness and accuracy, with a root mean square error of 0.48 kilograms per hour, whereas artificial neural networks tended to overfit and exhibited reduced performance in extended deployment periods. This finding underscores the importance of selecting models that combine predictive power with long term stability for environmental monitoring applications.

Although machine learning methods show great promise in improving scalability, adaptability, and predictive accuracy in environmental science, several challenges persist. Integrating disparate datasets such as point source emissions inventories, satellite observations, and meteorological grids requires sophisticated data fusion techniques to ensure consistency and reliability. Regulatory concerns about model transparency and data privacy also highlight the importance of developing explainable artificial intelligence systems for environmental monitoring. Addressing these challenges will be crucial for building trust in machine learning based frameworks and supporting their adoption in policy and operational contexts.

In summary, machine learning techniques ranging from tree-based ensemble models to advanced deep learning architectures offer powerful tools for greenhouse gas prediction and monitoring. They enable more responsive and fine-grained analyses of emissions patterns, which can inform more effective mitigation strategies. However, the realization of their full potential depends on overcoming barriers related to data quality, model explainability, and the ability to scale systems for continuous use in complex urban and regional environments.

2.3 Multi-Source Data Integration and Remote Sensing

Integrating data from multiple sources is increasingly recognised as a key strategy for enhancing methane monitoring systems. Relying on a single type of dataset, whether from ground-based sensors or satellite observations, often creates gaps in spatial and temporal coverage that can lead to inaccuracies in emission quantification. By combining diverse datasets, researchers can overcome these limitations and develop more complete and nuanced emissions inventories that better capture the dynamic nature of methane emissions. Multi-source integration also allows for cross-validation between independent measurement systems, which can enhance the reliability of monitoring frameworks and reduce uncertainty in emission estimates.

A systematic review highlights the core challenges associated with multi-source data integration, including calibration inconsistencies across instruments, temporal misalignments between datasets, and technical complexities when reconciling measurements with different spatial resolutions and accuracies. These challenges are especially pronounced when integrating data from satellites and ground-based sensors, as the differences in coverage and detection thresholds can result in significant discrepancies. Addressing these barriers requires the development of advanced fusion algorithms capable of harmonising heterogeneous datasets while maintaining scientific rigor and ensuring that uncertainties are properly quantified and propagated through the analysis process (Xu et al., 2025).

Atmospheric remote sensing has played a central role in expanding methane monitoring capabilities at regional and global scales. Satellite platforms provide a critical vantage point for detecting large-scale emission patterns and identifying hotspots that may not be apparent through ground-based monitoring alone. Recent studies have shown that integrating satellite data with airborne and ground-based measurements can substantially enhance detection accuracy and geographic coverage, particularly in areas where field measurements are sparse or logistically challenging to obtain. However, technical limitations such as cloud cover

interference, atmospheric scattering, and varying detection thresholds of different satellite instruments remain significant obstacles to achieving consistent and comprehensive global monitoring of methane emissions (Zhang et al., 2023).

Validation studies of satellite-based detection systems have demonstrated both the potential and current constraints of these technologies. One such study reported that existing satellite platforms can detect between 19 percent and 89 percent of methane emissions from oil and gas infrastructure, depending on environmental conditions, instrument sensitivity, and the scale of emission events (Sherwin et al., 2023). These findings highlight the strengths of satellite observations in identifying large emission sources but also underscore the need for hybrid systems that combine high-resolution satellite imagery with airborne surveys and ground-based networks. Such hybrid approaches have proven effective in improving the detection of transient super-emitter events, which are often highly variable and episodic in nature (Sherwin et al., 2023).

Machine learning techniques are playing an increasingly important role in supporting the integration of multi-source datasets. Advanced methods such as Gaussian Process Machine Learning, combined with probabilistic programming, have shown promise in analysing satellite data and improving emission estimates by capturing spatiotemporal covariance structures and reducing uncertainties. These approaches provide a flexible framework for modelling complex relationships within heterogeneous datasets and can facilitate the integration of information from diverse platforms in methane monitoring systems (Jeong et al., 2025).

In Canada, a hybrid multi-source fusion framework was developed that combined satellite, aerial, and ground-level data within an AI-driven system. This system demonstrated improved localisation and quantification of methane emissions compared to traditional monitoring approaches. By leveraging artificial intelligence techniques, the framework was able to synthesise large volumes of data from different sources and provide actionable insights for policymakers and environmental managers. This system also enhanced operational decision-making for emission mitigation strategies, particularly in complex urban environments where methane sources are diffuse and difficult to characterise using any single measurement technique (Yazdinejad et al., 2025).

Together, these studies highlight the critical role of multi-source data integration in advancing methane monitoring systems. Effective frameworks must balance trade-offs between spatial resolution, geographic coverage, and computational demands while leveraging advanced machine learning techniques capable of processing large and heterogeneous datasets. Addressing these technical and operational challenges will be essential to supporting robust, real-time methane monitoring that can inform climate change mitigation policies and interventions at both local and global scales.

2.4 Research Gaps and Contribution

Despite the substantial progress in methane (CH_4) monitoring and predictive modelling, several critical gaps remain that limit the operational deployment of such technologies in the UK context. Existing approaches, while leveraging advanced machine learning (ML) pipelines and multi-source data integration, often suffer from coarse spatial resolution, lack of localized validation, and insufficient consideration of sector-specific emission dynamics. This section highlights these gaps and positions the present study as a novel contribution that addresses them.

Gaps in Fine-Grained Methane Prediction

One key limitation in current CH_4 modelling efforts is the reliance on global or continental-scale datasets, which are ill-suited to capturing fine-grained emissions dynamics at sub-national levels. For instance, Cutler et al. (2025) developed a Random Forest-based framework integrating TROPOMI satellite methane data with UK-specific soil and land-use datasets. While their approach demonstrated improved predictive accuracy (RMSE = 29.48 ppb), challenges such as substantial data gaps due to cloud cover and difficulties in resolving small point sources persisted. Furthermore, the model's integration of multiple environmental datasets highlighted a persistent issue in temporal and spatial alignment, particularly in cloud-prone regions such as the UK. These findings underscore the need for high-resolution, grid-based predictions that effectively fuse point-source emissions with climate data to enable targeted mitigation strategies.

Global Inventories and Underrepresented Sources

At the global scale, methane inventories remain incomplete and often overlook significant contributors such as abandoned oil and gas wells. Lei et al. (2025)

produced a comprehensive inventory of methane emissions from 4.5 million abandoned wells worldwide, estimating emissions of 0.4 Mt in 2022. Notably, 60% of these emissions originated from unplugged wells, highlighting substantial uncertainties in national greenhouse gas reporting. Although such inventories provide valuable insights, they are less applicable to the UK context, where regulatory environments and emission patterns differ. The lack of localized, sector-specific studies incorporating abandoned wells, landfills, and agricultural activities within predictive models represents a critical gap in supporting effective UK methane mitigation policies.

Challenges in Operationalizing ML for Environmental Monitoring

Another significant gap lies in the operational deployment of ML models for real-time or near-real-time environmental monitoring. While ML techniques such as XGBoost and Random Forest have shown promise in other sustainability applications (Uppalapati et al., 2025), their adoption for live methane monitoring remains limited. Issues such as model explainability, data latency, and integration with heterogeneous datasets hinder their practical use in environmental decision-making contexts. As Uppalapati et al. (2025) demonstrated in the context of biochar yield prediction, incorporating explainable AI (e.g., SHAP analysis) enhances model transparency and stakeholder trust, a consideration equally critical for CH_4 monitoring systems.

The Novel Contribution of This Study

This dissertation addresses these gaps through the development of a fine-grained, grid-based CH_4 prediction model that integrates point-source emissions data with high-resolution climate variables from NetCDF datasets. By employing advanced ML techniques (Random Forest, XGBoost) and feature engineering strategies that account for spatial-temporal dynamics, the model provides a novel solution to the challenge of localized methane forecasting in the UK. Furthermore, the use of explainability tools enhances the operational potential of the model within GreentecAI's framework, facilitating actionable insights for policymakers and environmental planners. The project's focus on UK-specific emissions and its operational orientation differentiates it from prior studies, offering both academic and practical contributions to the field of environmental data science.

3. Methodology

This chapter outlines the approach used to build a predictive model for methane (CH_4) emissions across UK facilities. It covers data acquisition, preprocessing, feature engineering, model training, and evaluation. The methodology integrates spatial, sectoral, and climate data using machine learning and interpretable modeling techniques, designed to support GreentecAI's operational goals and broader environmental forecasting needs.

3.1 Data Collection

The dataset construction required integrating multiple data sources to capture the environmental, spatial, and sectoral factors influencing facility-level methane (CH_4) emissions across the UK. The approach combines point-source emissions records with high-resolution environmental data and contextual metadata, creating a robust foundation for predictive modelling.

This integration addresses a critical limitation in existing CH_4 modelling approaches, which often rely on single-source data and overlook the complex interactions between environmental conditions, industrial activity, and spatial clustering that drive emission variability. By systematically combining emissions data with multi-dimensional contextual information, this dataset enables the development of models capable of identifying actionable patterns for environmental monitoring and policy design.

The data are sourced as follows:

● **Emissions Data** comprises annual CH_4 emission records from the UK National Atmospheric Emissions Inventory (NAEI), covering 18,130 facility-year observations from 2018 to 2022. Each entry includes emission quantities (kilotons), spatial coordinates (British National Grid Easting and Northing), sector classifications using standardised industry codes, and facility identifiers. The dataset spans 25 industrial sectors, including Oil & Gas Exploration, Waste Collection and Treatment, Major Power Producers, and Natural Gas Processing & Distribution, supporting sectoral analysis of emissions.

● **Climate Data** was obtained from gridded NetCDF files provided by the UK Met Office, offering 1 km² spatial resolution across the UK. Monthly measurements include mean temperature (°C), total rainfall (mm), wind speed (m/s), and atmospheric pressure (hPa) for 2018–2022. Facility coordinates were transformed and spatially joined to the nearest grid cell, enabling extraction of site-specific environmental variables for seasonal and annual aggregation.

● **Spatial and Sectoral Metadata** includes regional classifications (England, Scotland, Wales, Northern Ireland), urban–rural designations, and facility density metrics calculated within 10 km and 25 km buffers around each point source. Sectoral context was provided via national reporting frameworks and regulatory classifications.

The final dataset maintains consistent five-year temporal coverage with complete alignment between emissions records and climate observations. This structure provides a robust foundation for temporal validation, trend analysis, and interpretable machine learning.

3.2 Data Preprocessing

To prepare the integrated dataset for modelling, a series of preprocessing steps were applied to ensure completeness, consistency, and compatibility with the machine learning pipeline. These included handling missing values, transforming the target variable, encoding categorical features, and scaling numerical variables where appropriate.

Methane emissions exhibited a pronounced right-skewed distribution, with a concentration of low-emitting facilities and a long tail of high-emission outliers. A logarithmic transformation was applied to the raw emission values to reduce skewness, stabilise variance, and enhance model interpretability ,a well-established practice in environmental modelling to normalise long-tailed distributions (West, 2021).

Missing values were primarily confined to climate features, arising from spatial mismatches during grid-cell assignment. Due to the low proportion of missing

entries, a simple imputation strategy using column-wise means was adopted. This approach is supported by recent work on the effectiveness of imputation techniques in structured datasets with minimal missingness (Tolou Shadbahr et al., 2023).

Facility locations, originally recorded in British National Grid format, were converted to latitude and longitude to enable alignment with the 1 km² NetCDF climate data. Once matched to their corresponding grid cells, original coordinate columns were removed to avoid redundancy.

Categorical attributes were processed to support machine learning compatibility. Sector classifications and regional labels were one-hot encoded, enabling the models to learn from discrete groupings without imposing artificial order. One-hot encoding is preferred over ordinal or label encoding for nominal variables due to its ability to eliminate spurious ordinality and improve algorithmic compatibility (GeeksforGeeks, 2019).

Standardisation was selectively applied to continuous features depending on model requirements. While tree-based models such as XGBoost and Random Forest are insensitive to feature scaling, linear models and neural networks benefit from zero-mean, unit-variance inputs. Standardisation improves numerical stability and training efficiency in gradient-based algorithms (Shaibu, 2024).

The resulting dataset was clean, numerically stable, and suitable for high-performance modelling. It served as the foundation for downstream feature engineering and model development described in the following sections.

3.3 Feature Engineering

Feature engineering was a critical step in capturing the diverse physical, spatial, and industrial patterns influencing methane (CH_4) emissions across UK facilities. This process involved aggregating seasonal climate indicators, computing spatial density metrics, generating climate interaction terms, and encoding industrial sectors. Each transformation was informed by domain expertise and structured to support both model interpretability and performance.

Climate Aggregates

High-resolution gridded climate data were aggregated to seasonal and annual metrics for each facility's geographic location. For variables such as temperature, wind speed, and atmospheric pressure, features were calculated for all four seasons (e.g., mean_temperature_summer, mean_pressure_winter). Rainfall, however, was aggregated annually (total_rainfall_annual) due to its higher spatial and temporal volatility and the need to smooth extreme values. This type of seasonal aggregation has been used effectively in environmental forecasting and energy-relevant climate predictions (Cionni et al., 2022). In total, 35 climate-based features were derived, enabling the models to incorporate both temporal and environmental variability.

Climate Interaction Terms

To capture non-linear interactions between climate variables that may influence methane release, 10 multiplicative interaction terms were created. These included both intra-seasonal and cross-variable combinations, such as mean_temperature_winter_X_total_rainfall_winter and mean_pressure_annual_X_mean_wind_annual. Interaction terms are increasingly used in environmental modelling to reflect complex dependencies between meteorological and chemical drivers (Li et al., 2025). These features were retained after evaluation for multicollinearity and signal contribution.

Spatial Features

Spatial characteristics were engineered to reflect each facility's geographic context and the potential for emission clustering. The original British National Grid coordinates were converted to latitude and longitude, then used to generate:

- Facility_Count_10km and Facility_Count_25km: counts of other facilities within defined radii
- UK_Region: macro-regional classification (e.g., England, Scotland, Wales)

- Urban_Rural_Class: binary urbanicity indicator

These features played a key role in highlighting spatial emission patterns and were among the top-ranked predictors in the SHAP analysis.

Sectoral Features

As introduced in the preprocessing stage (Section 3.2), the emissions dataset included a Sector field that categorised each facility into one of 25 industry groups, including Oil & Gas Exploration, Waste Treatment, and Major Power Producers. During feature engineering, this was prepared for one-hot encoding, enabling the models to learn sector-specific patterns. These features were found to be highly predictive of methane emissions.

Feature Summary and Visual Preview

The final engineered dataset (point_sources_climate) included 53 features prior to encoding, distributed across:

- 35 climate aggregates
- 10 climate interaction terms
- 6 spatial descriptors
- 2 categorical placeholders: Sector and Emitter_Category

While Sector was passed into the model post-encoding, Emitter_Category was used only for exploratory analysis to compare climate features between high and low emitters and was excluded from final modelling.

To maintain interpretability, no dimensionality reduction techniques (e.g., PCA) were applied. All transformations were guided by explainability objectives. **Figure 1** presents a representative snapshot of the final dataset grouped by feature type, including emission values (Emission, log_emission) used as modelling targets.

Figure 1. Final Feature Set After Feature Engineering

A sample of the dataset used in modelling, including climate, spatial, sectoral, and emission variables. Features are grouped and color-coded by transformation type to reflect the engineered structure prior to encoding.

Spatial Features

Feature Name	Example Value
Longitude	-1.491468612369704
Latitude	52.39595469652897
Facility_Count_10km	49
Facility_Count_25km	252
UK_Region	Midlands
Urban_Rural_Class	1

Climate Aggregates

Feature Name	Example Value
mean_temperature_annual	11.020678229891816
std_temperature_annual	5.231692011982668
mean_temperature_winter	5.076611686181453
std_temperature_winter	1.582552988739846
mean_temperature_spring	9.807074674025015
std_temperature_spring	3.5675551751399524
mean_temperature_summer	18.04083707685207
std_temperature_summer	1.3486520443246068
mean_temperature_autumn	11.15818948250873
std_temperature_autumn	2.515018639782768
total_rainfall_annual	605.9638657205028
total_rainfall_winter	173.11035517482577
total_rainfall_spring	226.50588500293614
total_rainfall_summer	86.28676002295185
total_rainfall_autumn	120.06086551978899
mean_pressure_annual	1014.5143900846724
std_pressure_annual	5.091005858341604
mean_pressure_winter	1014.1263275875725
std_pressure_winter	3.2480885667905826
mean_pressure_spring	1010.0931184373541
std_pressure_spring	6.66790655998437
mean_pressure_summer	1017.4211501060172
std_pressure_summer	1.5413856391533833
mean_pressure_autumn	1016.4169642077453
std_pressure_autumn	3.8104013021171212
mean_wind_annual	2.6250468209414155
std_wind_annual	0.3341050350372736
mean_wind_winter	2.8135382895744114
std_wind_winter	0.26415289660462055
mean_wind_spring	2.8052354366504417
std_wind_spring	0.32439693236064565
mean_wind_summer	2.2907785822398243
std_wind_summer	0.18482965052706746
mean_wind_autumn	2.590634975300983
std_wind_autumn	0.23752089175872806

Emission Targets

Feature Name	Example Value
Emission	65.97
log_emission	4.204244757921959

Climate Interaction Terms

Feature Name	Example Value
mean_temperature_annual_X_total_rainfall_annual	6678.132783047033
mean_temperature_winter_X_total_rainfall_winter	878.8140520795424
mean_temperature_spring_X_total_rainfall_spring	2221.3601283299176
mean_temperature_summer_X_total_rainfall_summer	1556.685379463507
mean_temperature_autumn_X_total_rainfall_autumn	1339.6618869038045
mean_pressure_annual_X_mean_wind_annual	2663.1477744910885
mean_pressure_winter_X_mean_wind_winter	2853.2832531331183
mean_pressure_spring_X_mean_wind_spring	2833.5490101572173
mean_pressure_summer_X_mean_wind_summer	2330.6865797806736
mean_pressure_autumn_X_mean_wind_autumn	2633.1653369658325

3.4 Model Selection and Training

This section presents the machine learning strategy developed to estimate methane (CH_4) emissions at the facility level using spatial, climatic, and industrial features. The process involved selecting appropriate algorithms, constructing training pipelines, applying temporal validation, performing hyperparameter tuning, and preparing final models for comparison. Modelling decisions were made to balance predictive performance, interpretability, and scalability, while maintaining methodological rigour consistent with environmental forecasting practices.

3.4.1 Model Selection Strategy

Three regression models were selected to benchmark performance across different levels of complexity and interpretability:

- **Linear Regression (baseline):** A regularised linear model was included as a transparent baseline for understanding how much predictive performance could be gained through more complex models. Although linear models are limited in their capacity to capture non-linear relationships, they provide a reference point for assessing the added value of tree-based approaches.

- **Random Forest Regressor:** This ensemble method was chosen for its ability to capture non-linearities and interactions between variables without heavy preprocessing. It is widely used in environmental prediction contexts and offers robustness to overfitting and multicollinearity.

- **XGBoost Regressor:** Gradient boosting was selected for its state-of-the-art performance on structured tabular data and its compatibility with SHAP for interpretability. XGBoost has been successfully applied in various emissions-related modelling tasks, including carbon and air pollution forecasting (Nebojša Balać et al., 2025; Alam et al., 2025).

To test trade-offs between performance and interpretability, each model was trained under two input configurations:

1. A **full feature model**, using all ~80 engineered and encoded features

2. A **strategic feature model**,using a curated subset of 14 features selected through exploratory data analysis (EDA), SHAP importance ranking, and domain-specific reasoning

This dual-track modelling framework allowed for comparison between complexity-driven and parsimonious approaches.

3.4.2 Temporal Data Splitting and Training Design

To simulate real-world deployment conditions and prevent information leakage, the dataset was partitioned chronologically based on the Year variable. Observations from **2018 to 2021** were used for training and validation, while **2022** was reserved as a holdout test set for final model evaluation. This 70/15/15 split ensured temporal integrity, an essential factor in emissions forecasting where patterns shift across years due to climate, regulation, and economic conditions (Chai and Draxler, 2014).

All preprocessing steps including log transformation of emissions, one-hot encoding of categorical variables, and imputation of missing values were embedded within the modelling workflow to preserve consistency and avoid data leakage. The target variable, CH_4 emissions, was log-transformed based on its highly skewed distribution, which is typical in environmental datasets with a small number of high-emitting facilities.

Two distinct pipelines were developed:

- One using **all features** including full climate aggregates, interaction terms, and categorical encodings
- One using the **strategic feature set** focused on the most informative and operationally practical predictors

3.4.3 Hyperparameter Optimisation

Hyperparameter tuning was applied specifically to the XGBoost models using **Optuna**, a modern hyperparameter optimisation framework that employs Bayesian techniques to efficiently search the parameter space (Akiba et al., 2019). The tuning process aimed to minimise validation error on 2021 data while avoiding overfitting through regularisation and sampling strategies.

The key hyperparameters tuned included:

- Number of estimators (trees)
- Learning rate
- Maximum tree depth
- Subsample and column sampling ratios
- L1 and L2 regularisation terms

Separate optimisation routines were run for the **full** and **strategic** XGBoost pipelines, allowing each to operate under its most favourable configuration. Other models (Random Forest and Linear Regression) were retained with minimal tuning, as they were included primarily for benchmarking and interpretability rather than performance maximisation.

3.4.4 Evaluation Metrics and Final Model Preparation

Three core regression metrics were selected to evaluate model performance:

- **R^2 Score**: Measures the proportion of variance explained by the model
- **Mean Absolute Error (MAE)**: Captures the average absolute deviation from observed values
- **Root Mean Squared Error (RMSE)**: Penalises larger errors more heavily and is particularly useful when large deviations are costly or policy-relevant (Chai and Draxler, 2014)

These metrics provide a balanced view of model generalisation, interpretability, and sensitivity to extreme values. They will be reported and interpreted in Chapter 4.

Following training and validation, the best-performing model, the Optuna-tuned XGBoost using the strategic feature set was retrained on the full 2018–2021 dataset. This final version was exported and prepared for integration into downstream tools such as the interactive dashboard and SHAP-based interpretability analysis.

4. Analysis and Findings

4.1 Exploratory Data Analysis: Spatial and Sectoral Patterns

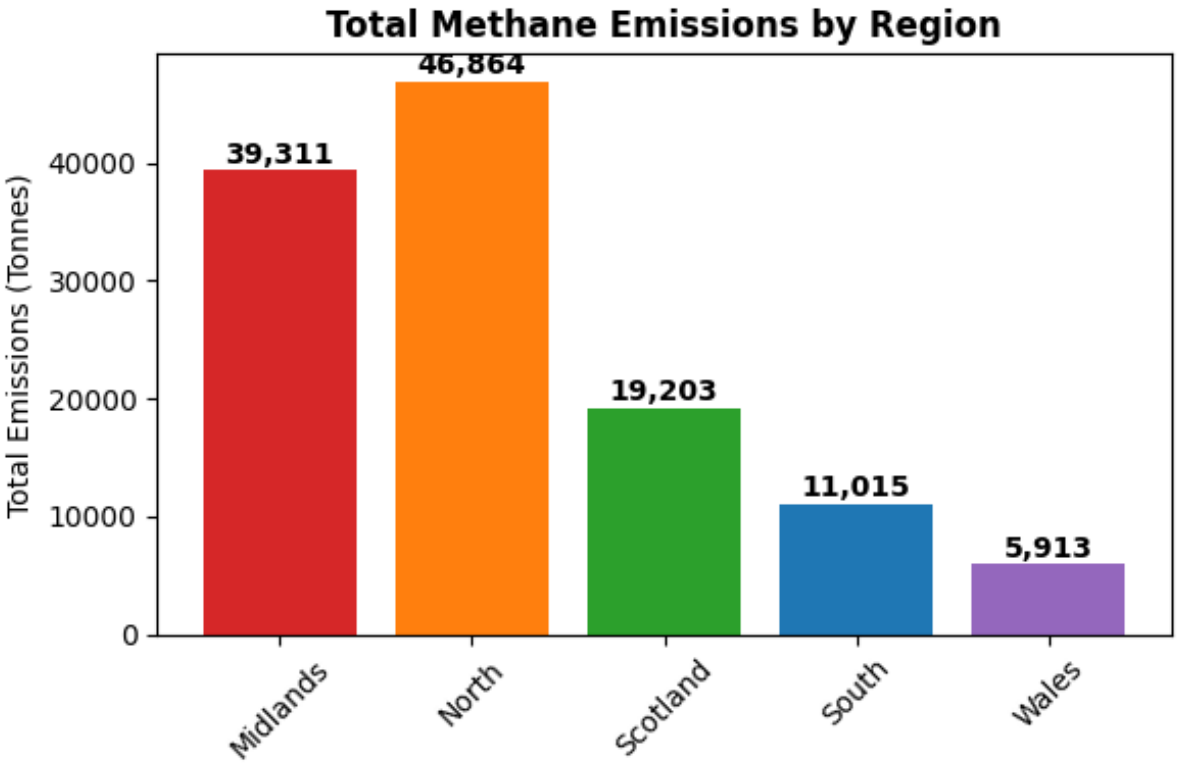
This section explores the spatial, sectoral, and climatic characteristics of methane (CH_4) emissions across UK facilities from 2018 to 2022. Exploratory Data Analysis (EDA) was used to identify regional clustering, industrial hotspots, and environmental variability, providing foundational insight for the predictive modelling strategy introduced later in this chapter.

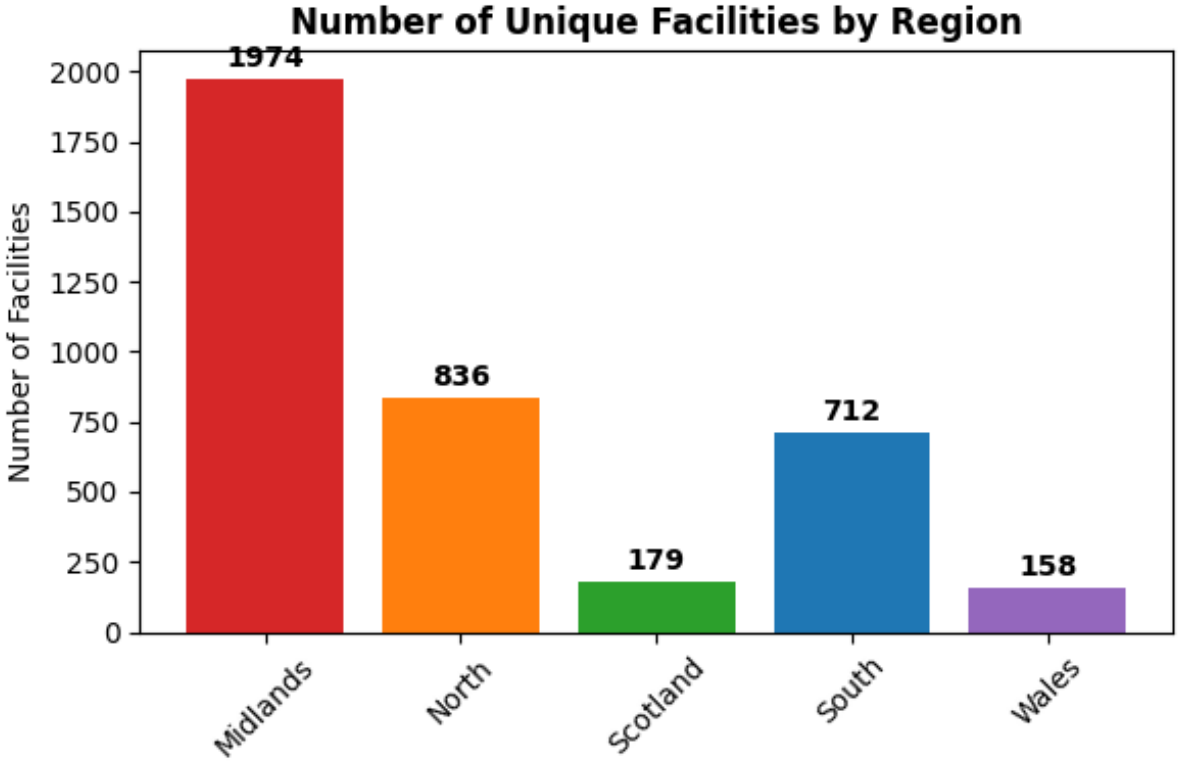
4.1.1 Regional Distribution of Emissions

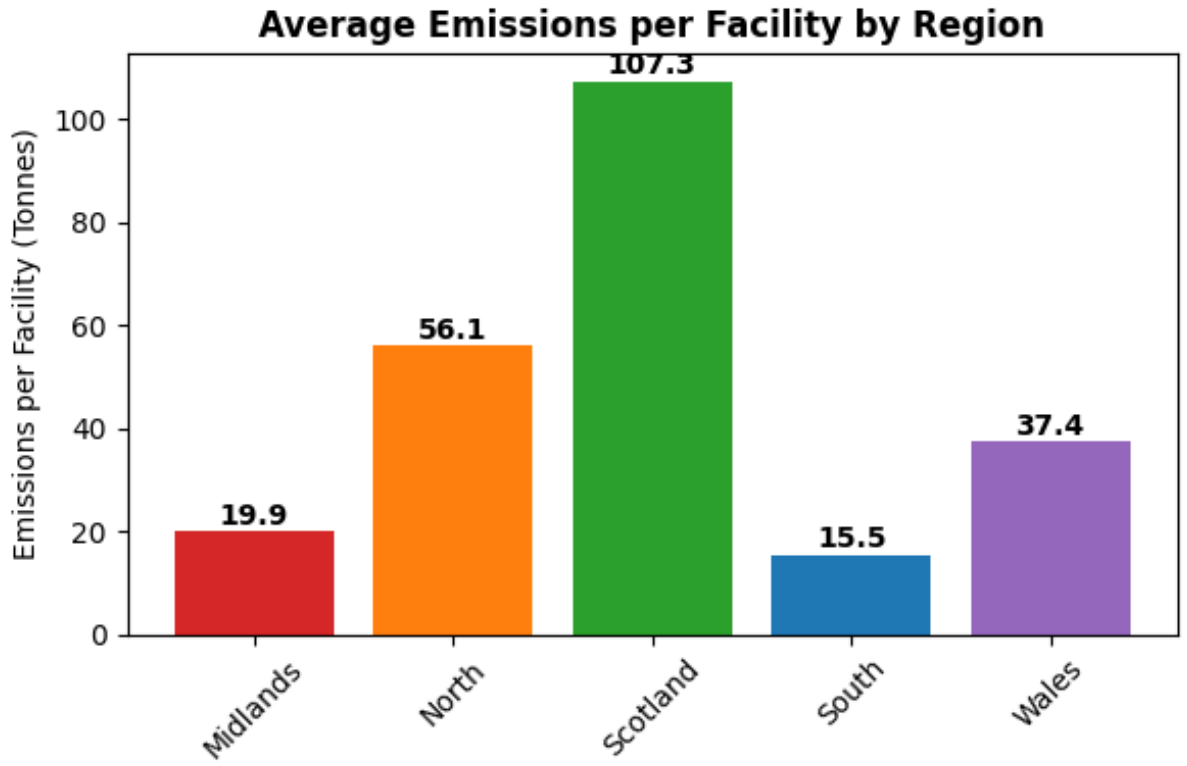
Methane emissions vary widely across UK regions, both in absolute volume and in per-facility intensity. As shown in **Figure 2**, the North and Midlands account for the highest total emissions, reflecting the concentration of heavy industrial infrastructure in these areas. In contrast, Scotland demonstrates the highest average emissions per facility, despite having a lower total number of emitters. This suggests the presence of disproportionately large facilities, such as oil and gas processing plants, in more remote regions.

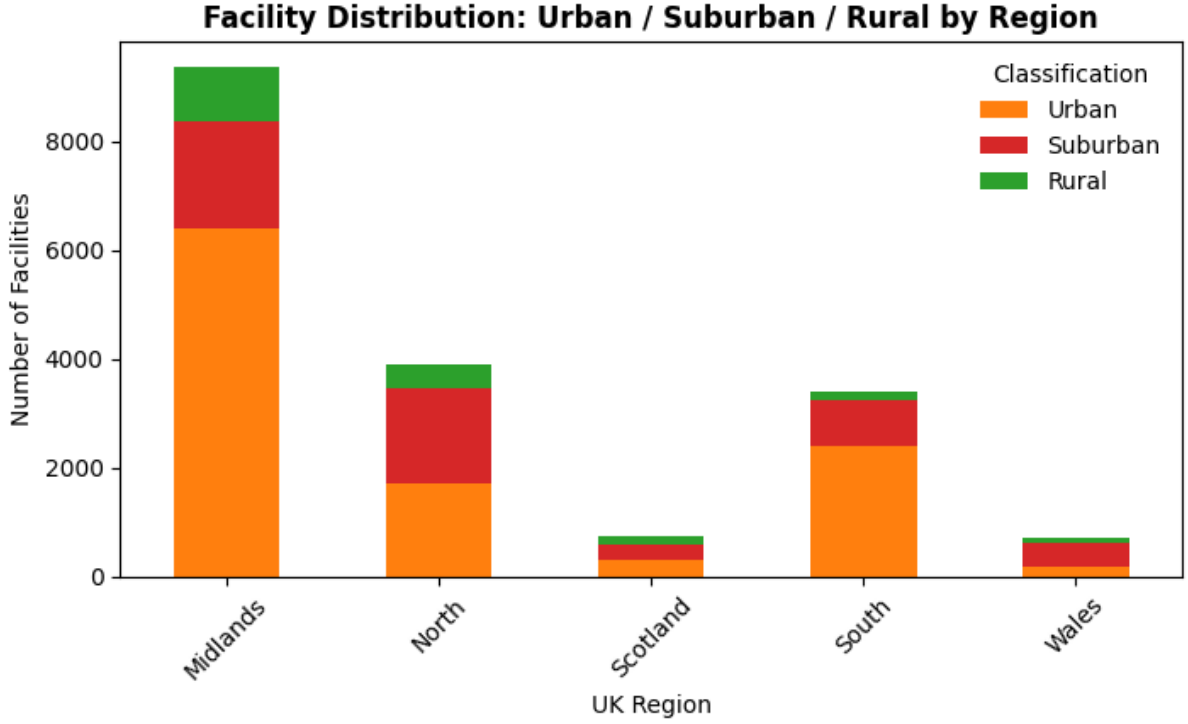
In addition, urban–rural classification highlights that urban and suburban facilities dominate CH_4 emissions output, aligning with expectations of higher industrial density. These findings support the inclusion of geographic, urbanicity, and facility density indicators as spatial features in the model.

Figure 2. UK Regional Methane Emissions Analysis (2018–2022)
Total emissions, facility counts, average emissions per site, and urban–rural classification for UK macro-regions based on point-source facility data.


Total Methane Emissions by Region
Total Emissions (Tonnes)
0
10000
20000
30000
40000
39,311
46,864
19,203
11,015
5,913
Midlands
North
Scotland
South
Wales


Number of Unique Facilities by Region
Number of Facilities
0
250
500
750
1000
1250
1500
1750
2000
1974
836
179
712
158
Midlands
North
Scotland
South
Wales


Average Emissions per Facility by Region
Emissions per Facility (Tonnes)
0
20
40
60
80
100
19.9
56.1
107.3
15.5
37.4
Midlands
North
Scotland
South
Wales

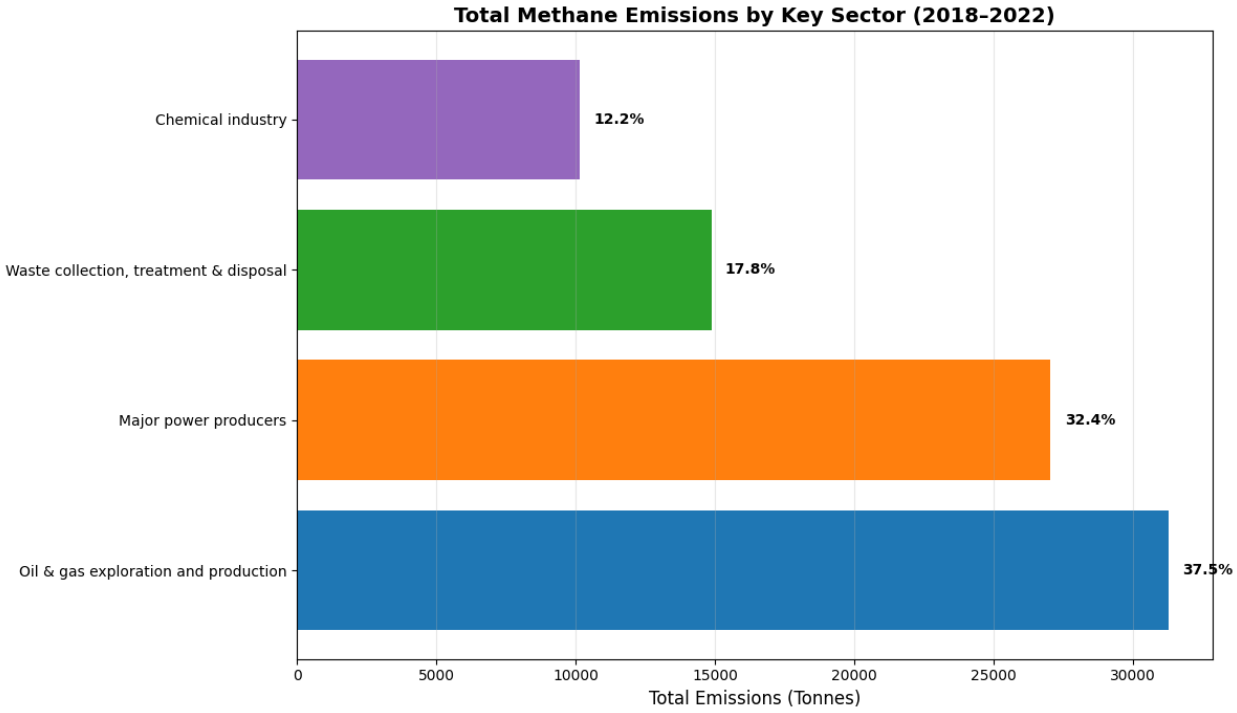

Facility Distribution: Urban / Suburban / Rural by Region
Number of Facilities
0
2000
4000
6000
8000
Classification
Urban
Suburban
Rural
Midlands
North
Scotland
South
Wales
UK Region

4.1.2 Sectoral Emissions Concentration

CH_4 emissions are highly skewed toward a small number of industrial sectors. As illustrated in **Figure 3**, four sectors , Oil and Gas Exploration, Major Power Producers, Waste Management, and the Chemical Industry contribute the majority of emissions nationwide. The Oil and Gas sector alone accounted for over one-third of all facility-level emissions during the five-year study period.

This sectoral skew justifies the one-hot encoding of industry classification and motivates future consideration of sector-specific model tuning or regulatory interventions.

Figure 3. Total Methane Emissions by Key Sector (2018–2022)
Horizontal bar chart showing total CH_4 emissions aggregated by top contributing industrial sectors.



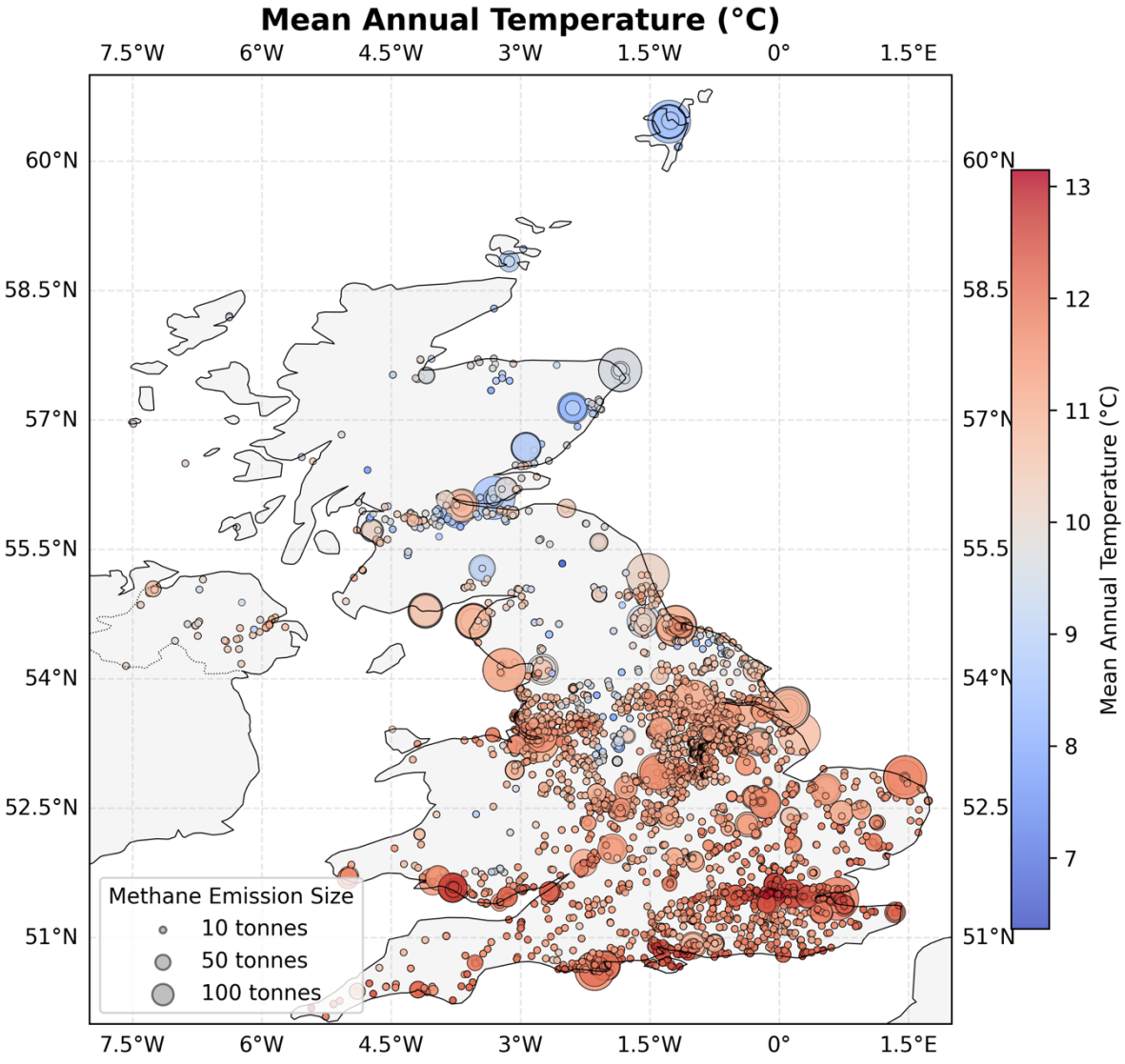
4.1.3 Climate–Emission Spatial Overlays

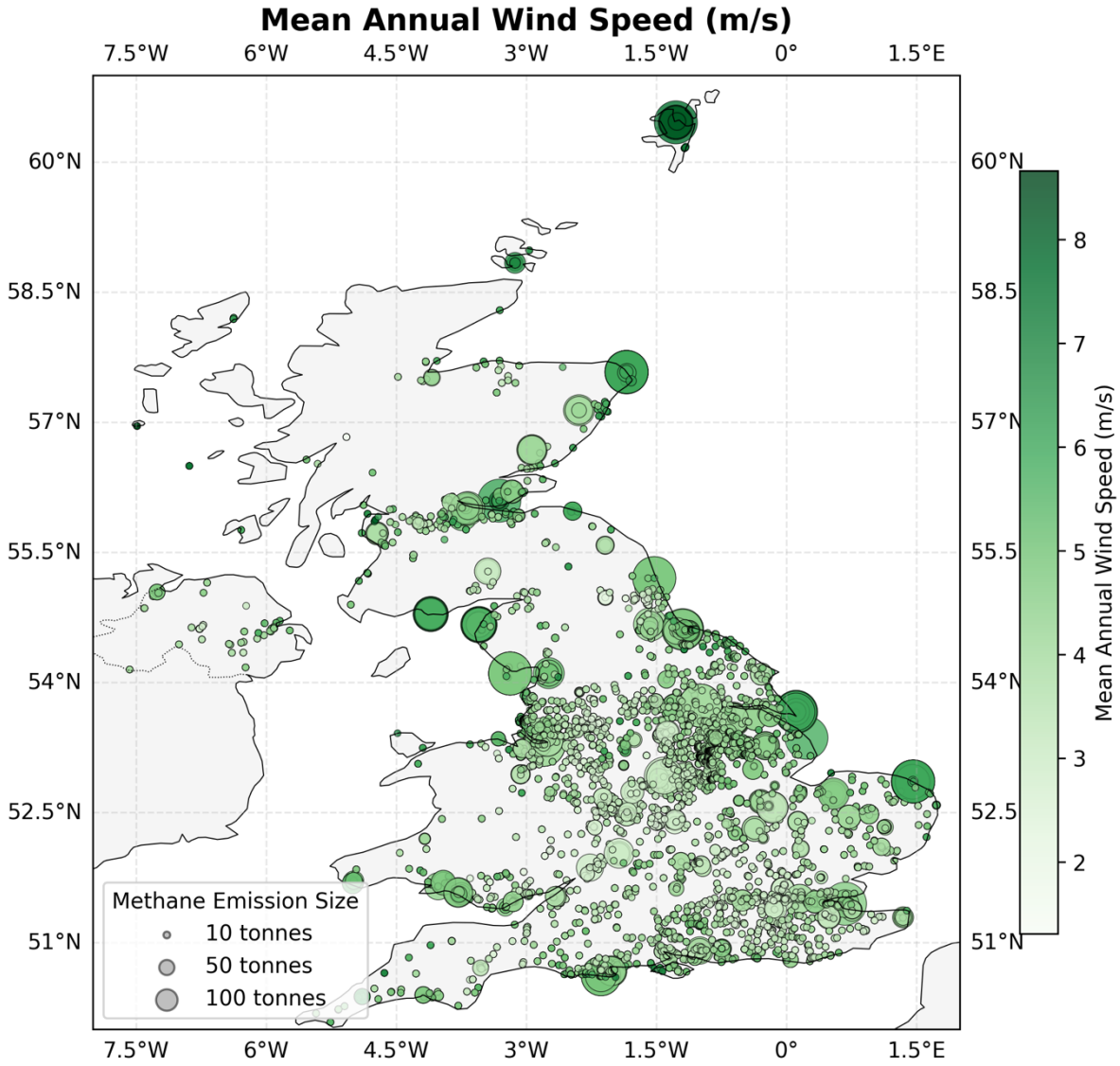
To better understand how emissions relate to environmental conditions, **Figure 4** visualises CH_4 emissions geographically, overlaid with four climate variables: temperature, rainfall, atmospheric pressure, and wind speed. Emission volume is

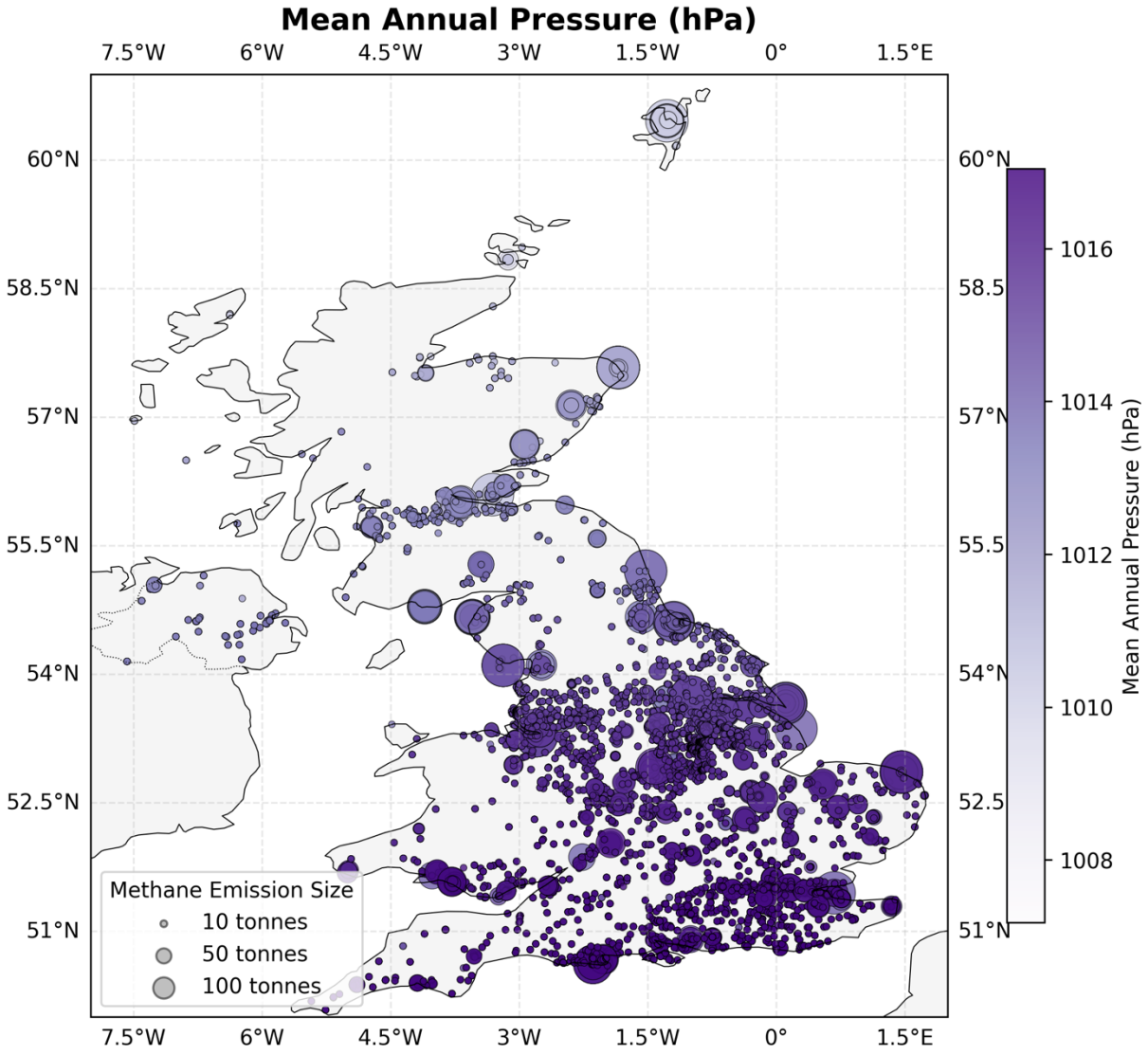
represented by bubble size, while climate intensity is depicted through colour gradients.

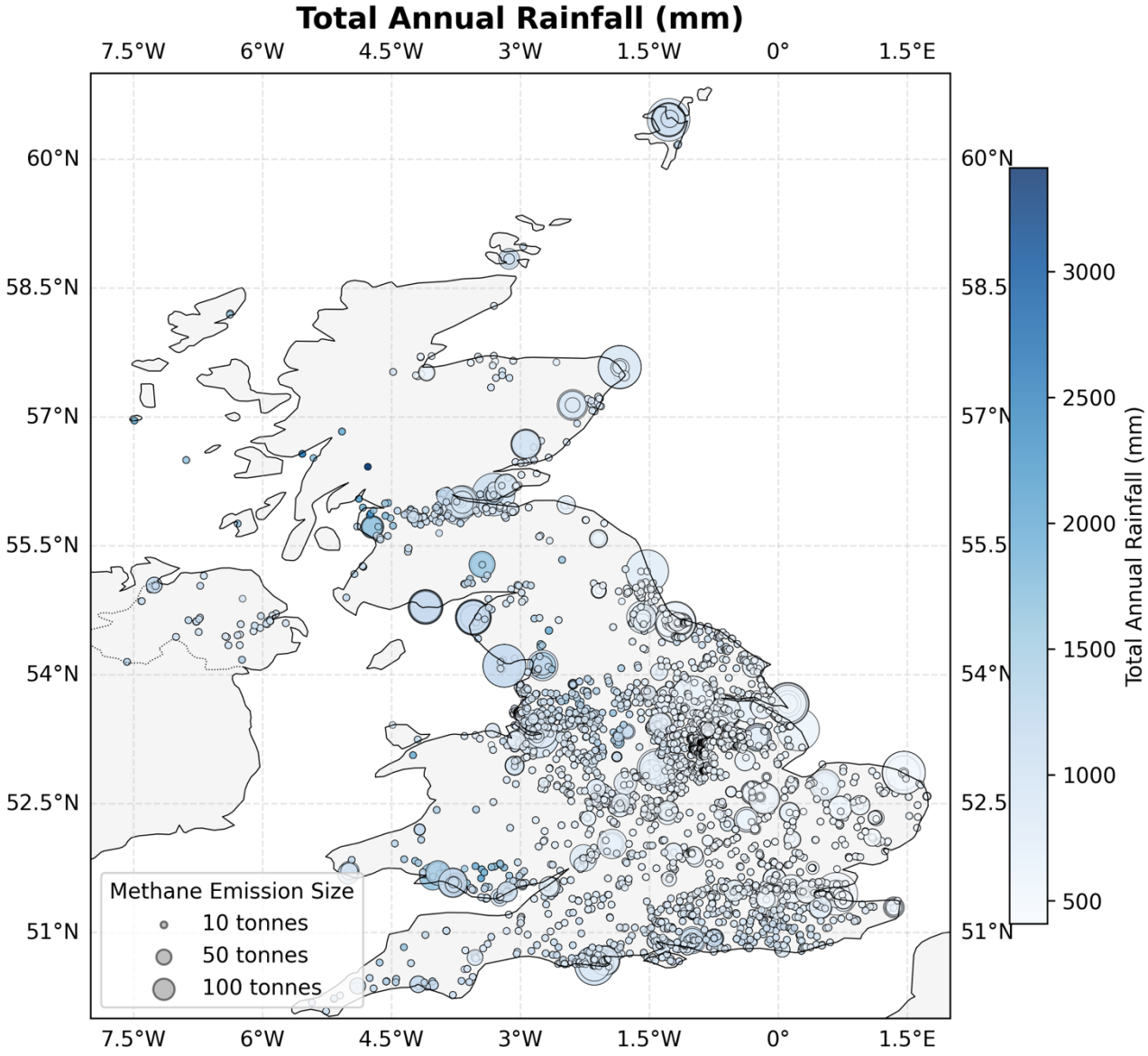
These overlays reveal that temperature and wind speed show stronger visual alignment with high-emission clusters, particularly in southern and eastern England, where dense industrial activity overlaps with warmer and windier conditions. Rainfall and pressure, by contrast, appear more diffuse and may influence emissions in less direct ways.

Figure 4. Spatial Distribution of Methane Emissions with Climate Overlays
Bubble maps showing facility-level CH_4 emissions, coloured by annual average temperature, rainfall, pressure, and wind speed across the UK.




Mean Annual Wind Speed (m/s)
7.5°W
6°W
4.5°W
3°W
1.5°W
0°
1.5°E
60°N
58.5°N
57°N
55.5°N
54°N
52.5°N
51°N
Methane Emission Size
10 tonnes
50 tonnes
100 tonnes
Mean Annual Wind Speed (m/s)
2
3
4
5
6
7
8


Mean Annual Pressure (hPa)
7.5°W
6°W
4.5°W
3°W
1.5°W
0°
1.5°E
60°N
58.5°N
57°N
55.5°N
54°N
52.5°N
51°N
Methane Emission Size
10 tonnes
50 tonnes
100 tonnes
Mean Annual Pressure (hPa)
1016
1014
1012
1010
1008


Total Annual Rainfall (mm)
7.5°W
6°W
4.5°W
3°W
1.5°W
0°
1.5°E
60°N
58.5°N
57°N
55.5°N
54°N
52.5°N
51°N
Methane Emission Size
10 tonnes
50 tonnes
100 tonnes
Total Annual Rainfall (mm)
3000
2500
2000
1500
1000
500

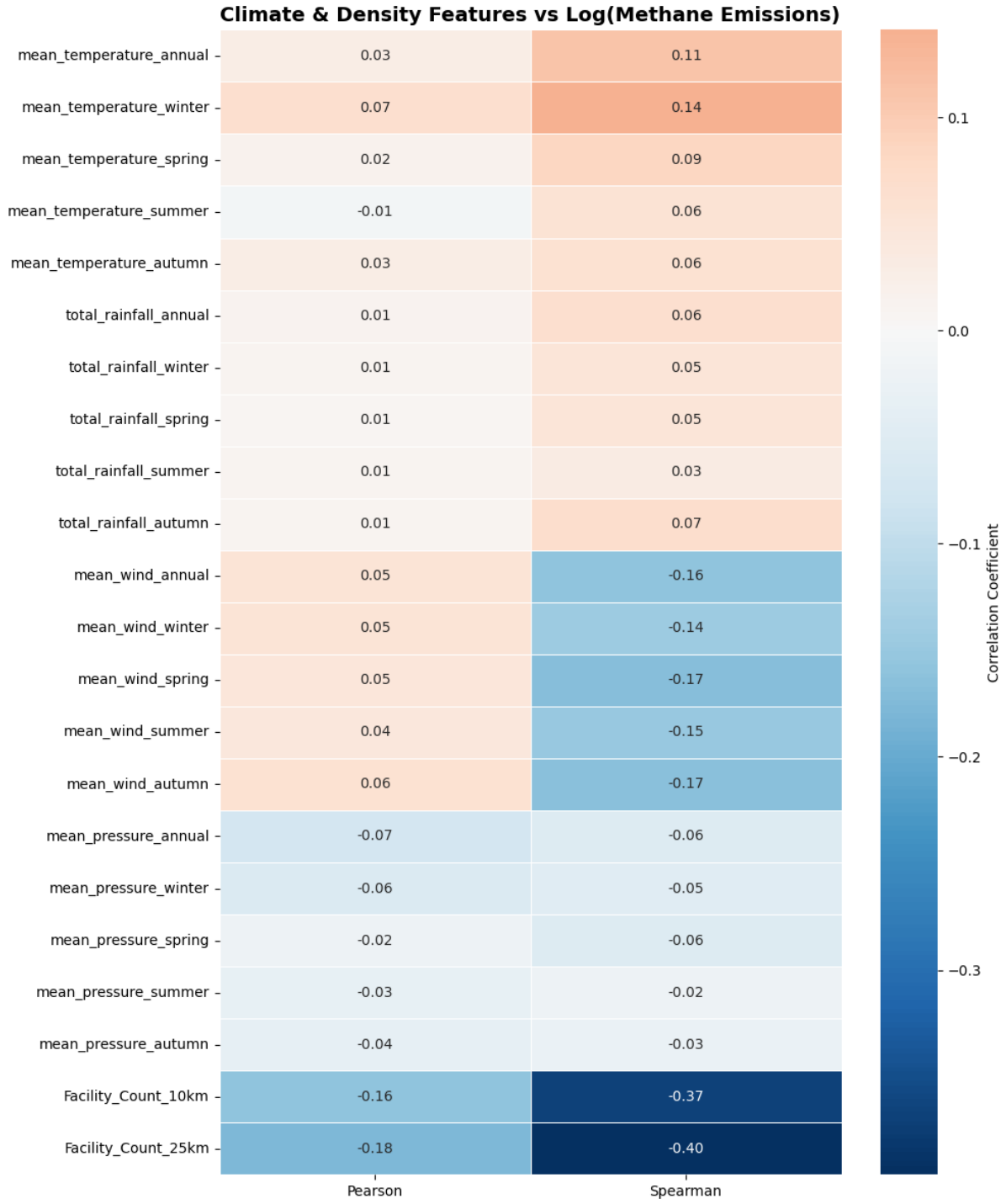
4.1. 4 Climate Correlation Matrix

To quantify the observed patterns, Figure 5 presents a correlation heatmap between climate, spatial, and emission-related features. Both Pearson and Spearman coefficients were computed to capture linear and monotonic relationships with the log-transformed emission target.

While absolute correlation magnitudes are relatively low, clear relative patterns emerge. Facility density within 10 km and 25 km buffers shows the strongest negative correlations with emissions, particularly under Spearman's metric (up to −0.40). Mean wind speed and temperature variables exhibit comparatively stronger relationships among the climate features, justifying their inclusion as standalone predictors. In contrast, rainfall and pressure show weak and inconsistent correlations, suggesting they are better represented through interaction terms rather than in isolation.

Figure 5. Correlation Matrix: Climate & Spatial Features vs Log (Emissions)

Heatmap showing pairwise Pearson and Spearman correlations between engineered predictors and the log-transformed methane emissions target.



4.1.5 Modelling Strategy Rationale

Based on the above findings, the modelling strategy prioritised**:**

- **Winter temperature and wind speed** as standalone climate variables, due to their relatively stronger spatial patterns and moderate statistical correlations

with emissions, especially when compared to other seasonal aggregates.

- **Rainfall and pressure** through interaction terms, as their influence on emissions appears more context-dependent and non-linear.

- **Sector and regional encodings**, reflecting industrial and geographic heterogeneity in methane output, as demonstrated in both SHAP and EDA findings.

- **Spatial clustering metrics**, **Facility_Count_25km**, which showed the highest negative correlation with log emissions, highlighting its value in modelling localised emission intensity.

- **Urban–rural classification**, to capture structural differences in emission dynamics across settlement types.

4.2 Model Evaluation and Comparison

This section presents the comparative evaluation of five supervised machine learning models trained to predict log-transformed CH_4 emissions. These models varied by algorithm and feature strategy and were assessed using standard regression metrics: **R^2**, **Mean Absolute Error (MAE)**, and **Root Mean Squared Error (RMSE)** on both validation (2021) and test (2022) sets.

Model Comparison

Figure 6 summarises performance across the five evaluated models:

- **XGBoost (All Features)**
- **Random Forest (All Features)**
- **XGBoost (Strategic Features)**

- **Random Forest (Strategic Features)**
- **XGBoost (Optuna-Tuned, Strategic Features)**

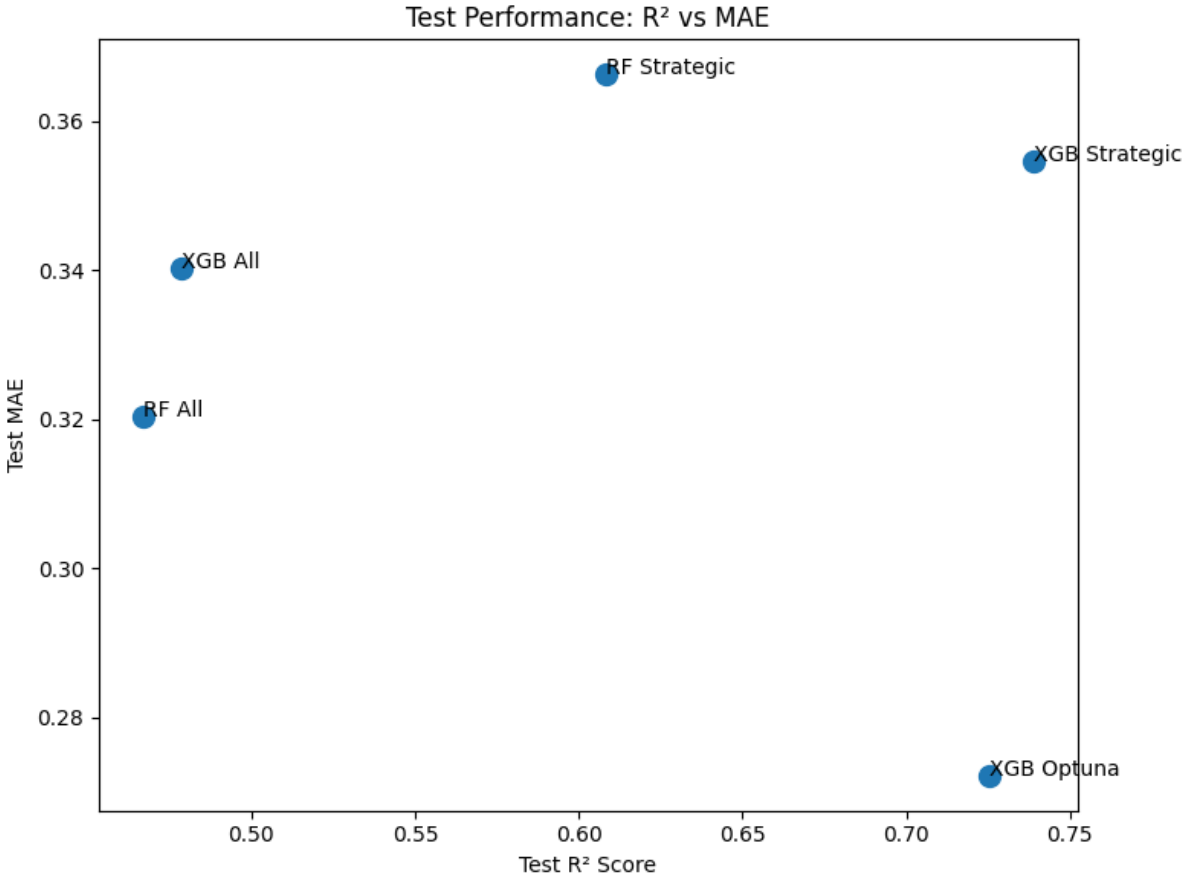
Among the baseline models, XGBoost consistently outperformed Random Forest across all metrics, particularly in generalisation. Both “All Feature” models exhibited signs of overfitting, with validation R^2 significantly higher than test R^2, most notably in the Random Forest (All) variant, which had a generalisation gap of 0.42.

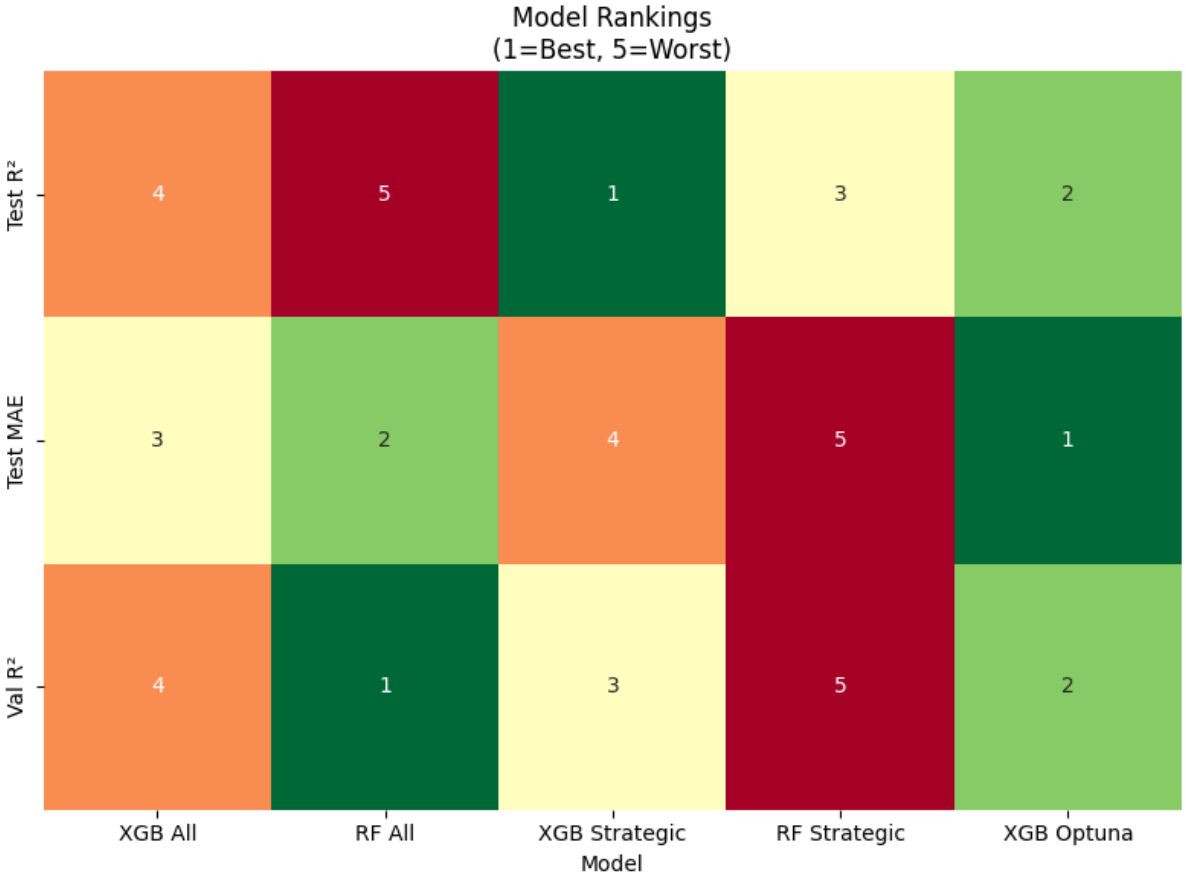
In contrast, strategic feature selection informed by SHAP importance, domain knowledge, and correlation analysis improved model parsimony and reduced overfitting. The **XGBoost (Strategic)** model achieved strong validation performance ($R^2 = 0.79$) with a low generalisation gap (0.04), suggesting better stability.

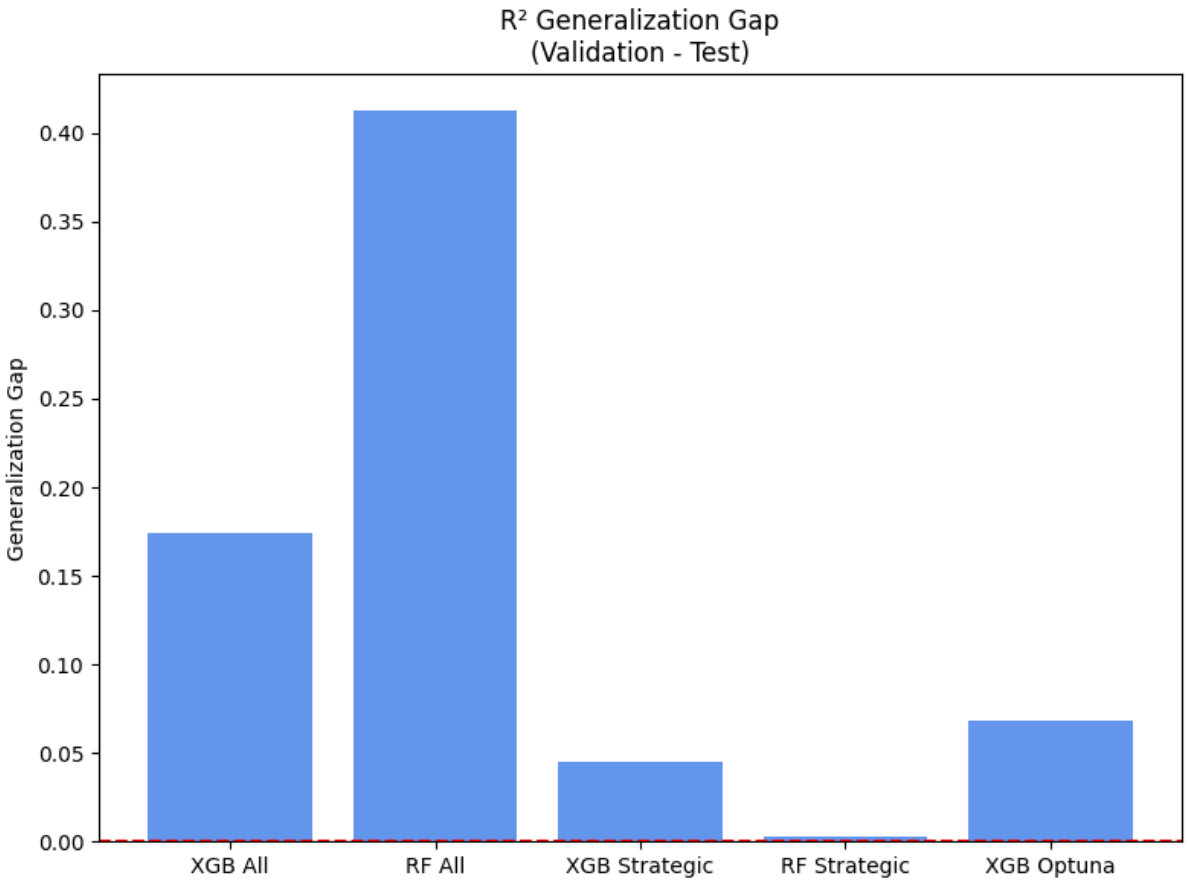
The final model, **XGBoost (Optuna)**, was trained on the 14 most important features using an optimized set of hyperparameters. It achieved the **highest test R^2 (0.72)** and **lowest test MAE (0.27)**, indicating superior predictive accuracy and robustness.

Figure 6. Model Performance Comparison: Methane Emission Prediction

Composite chart comparing R^2, MAE, generalisation gap, and test set performance for all five models. Also includes a ranking matrix (1=best, 5=worst) based on performance metrics.


Test Performance: R² vs MAE
RF Strategic
XGB Strategic
XGB All
RF All
XGB Optuna
0.36
0.34
0.32
0.30
0.28
Test MAE
0.50
0.55
0.60
0.65
0.70
0.75
Test R² Score


Model Rankings
(1=Best, 5=Worst)
Test R²
Test MAE
Val R²
4
5
1
3
2
3
2
4
5
1
4
1
3
5
2
XGB All
RF All
XGB Strategic
RF Strategic
XGB Optuna
Model

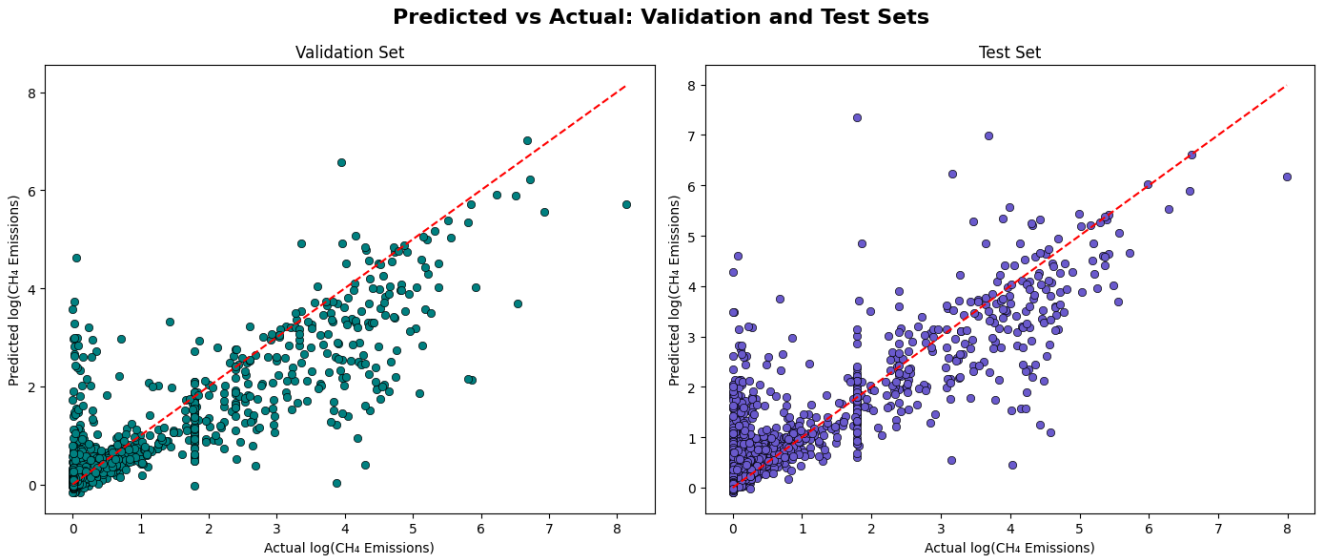


Final Model Fit

To assess how well the final model aligned with observed values, Figure 7 visualises **predicted vs actual log (CH_4) emissions** for both the validation and test sets. The tight clustering around the diagonal line reflects strong predictive alignment. Most deviations occur in high-emitting outliers, a known challenge in skewed environmental datasets.

Figure 7. Predicted vs Actual Emissions (Validation and Test Sets)

Scatter plots comparing model predictions to actual log-transformed CH_4 emissions. The dashed red line represents the ideal 1:1 fit, used to visually assess prediction accuracy.



4.3 SHAP-Based Interpretation and Insights

To interpret the final XGBoost model trained on the strategic feature set, SHAP (SHapley Additive exPlanations) was employed to quantify each variable's marginal impact on predicted methane emissions. SHAP provides a mathematically grounded approach to feature attribution, ensuring that even non-linear and interaction effects can be explained in a consistent and local-global framework.

Global Feature Importance

Figure 8 displays the SHAP summary plot, ranked by average absolute impact on model output. The top predictors highlight a blend of sectoral, spatial, and environmental drivers:

- **Sectoral indicators** dominated the top ranks, led by cat__Sector_Oil & gas exploration and production and cat__Sector_Major power producers, aligning with earlier domain expectations.

- **Spatial coordinates** — num__Latitude, num__Longitude, and num__Facility_Count_25km also exhibited strong influence, underscoring the geographic clustering of high-emission facilities.

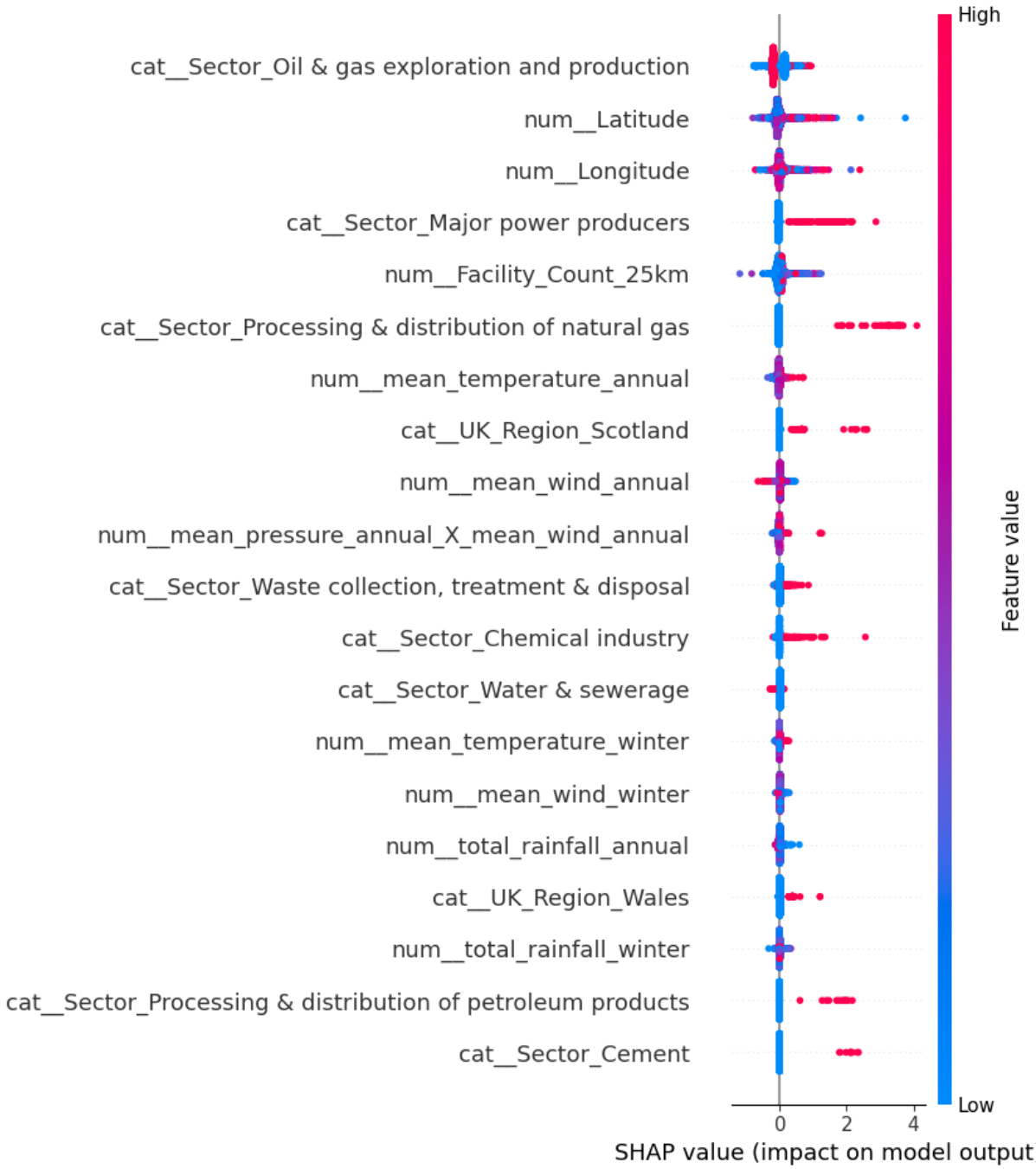
- **Climate variables** such as num__mean_temperature_annual and num__mean_wind_annual ranked within the top 10, validating their inclusion in the strategic feature set.

- Notably, the interaction feature num__mean_pressure_annual_X_mean_wind_annual had clear marginal impact despite low raw correlations — confirming its non-linear contribution to emission behavior.

These results validate the strategic down-selection of features and show that emission patterns are driven by a synergy of industrial, locational, and climatic factors.

Figure 8. SHAP Summary Plot: Global Feature Importance

Summary of feature influence using SHAP values. The x-axis shows the marginal impact on log-emission predictions, while colour denotes relative feature magnitude. Sectors, location, and select environmental variables show the strongest global importance.



In addition to global trends, SHAP values were used to explain individual predictions for high-emitting sites. Common patterns among the top 10 highest-prediction cases included:

- Proximity to other facilities (num__Facility_Count_25km)

- Classification under cat__Sector_Oil & gas exploration and production

- Location in cat__UK_Region_Scotland or high-latitude zones

- Elevated mean wind speed and winter temperature values

These compound factors created emission-prone profiles identifiable before emissions are observed, supporting preventative monitoring strategies.

5. Discussion

5.1 Model Interpretation via SHAP

To interpret the predictive behaviour of the final xgb_strategic_optuna model, SHAP (SHapley Additive Explanations) was employed to evaluate how each feature contributed to methane (CH_4) emission predictions. This method provides a robust, model-agnostic framework for assigning additive importance scores to individual features, helping unpack the “black box” behaviour often associated with ensemble models.

The analysis revealed that industrial sector encodings were the most influential predictors. Facilities associated with Oil & Gas exploration, Major Power Producers, and Natural Gas Processing consistently contributed to higher predicted emissions. These sectoral indicators reflect underlying process-related methane intensity and serve as reliable proxies for operational emissions risk.

Spatial characteristics, particularly latitude, longitude, and regional density metrics like Facility_Count_25km, were also ranked highly. This supports earlier spatial analysis (Section 4.1), suggesting that geographical clustering and regional context are key determinants of emission behaviour. The inclusion of UK_Region_Scotland as a top contributor further aligns with earlier findings highlighting disproportionate emissions in Scotland.

Climate variables such as mean annual temperature, winter temperature, and annual wind speed were among the top-ranked numerical features. These variables likely capture ambient conditions that modulate methane release and dispersion. Notably, an interaction term (mean_pressure_annual_X_mean_wind_annual) also featured prominently, indicating complex dependencies between atmospheric dynamics and emission levels.

These findings echo the results of Baba et al. (2025**)**, who applied XGBoost and SHAP to predict methane emissions at a cattle market. Their study identified temporal and seasonal features, such as wet-season climate patterns, as dominant drivers. In both cases, SHAP facilitated the extraction of environmental predictors with domain relevance.

Similarly, Sysoeva et al. (2024) used SHAP to disentangle the role of climate and facility-level factors in methane concentrations across Alberta's oil sands. Their model highlighted the dual role of atmospheric features (e.g., ozone) and facility configuration, reinforcing the idea that emissions stem from a fusion of environmental and industrial dynamics. Our findings mirror this, showing that the interaction of sectoral activity, spatial clustering, and seasonal climate variation yields the strongest predictive signal.

In summary, the SHAP-based interpretation confirms that methane emissions are governed by a multifactorial interplay of sector, location, and seasonal conditions. This reinforces the value of interpretable machine learning in identifying actionable levers for emission mitigation, such as prioritizing inspections at clustered industrial zones or adapting monitoring strategies to seasonal variation.

5.2 Business and Environmental Implications

The findings from this project present several insights with direct relevance for emissions reduction strategies and environmental analytics adoption in the energy and industrial sectors. While the models developed are not designed for direct deployment, the results highlight data-driven approaches that could inform GreentecAI's prototype refinement and broader environmental monitoring practices.

5.2.1 Sector-Specific Risk Profiling and Resource Allocation

One of the key takeaways from the strategic feature XGBoost model is the consistently high contribution of sectoral variables, particularly Oil & Gas Exploration and Major Power Producers to predicted CH_4 levels. These sectors not only appeared among the top SHAP-ranked features but also clustered spatially in high-emission regions. This enables organisations to adopt risk-based prioritisation frameworks, whereby monitoring resources are dynamically allocated to facilities and sectors associated with historical emission volatility.

For example, regulators or internal ESG auditors could weight inspection frequency and depth according to model-predicted risk levels. Waste treatment plants with seasonally fluctuating emissions or remote gas exploration sites near emission

thresholds may benefit from adaptive monitoring regimes, even if they currently meet standard reporting criteria.

5.2.2 Regulatory and Policy Applications

The integration of facility-level emissions data with climate and spatial metadata offers potential pathways for next-generation regulatory compliance frameworks. Policymakers and environmental agencies increasingly seek scalable tools for detecting anomalies, assessing sectoral compliance, and forecasting emission surges. The current approach demonstrates that ML models can provide actionable summaries, interpretable drivers, and even geographic insights that extend beyond traditional audit checklists.

Importantly, such models should be seen as decision-support tools rather than enforcement mechanisms. As highlighted in broader policy discussions, especially in the oil and gas industry, standardised methane monitoring protocols are critical for achieving international climate commitments. The application of AI-powered predictions in this domain is promising but must be paired with human-in-the-loop validation, contextual understanding, and rigorous ethical controls (Esiri and Ekemezie, 2024).

5.2.3 Real-Time Monitoring and Scalable Deployment

In operational contexts, the potential exists to use models like the one developed in this study as part of real-time dashboards for environmental compliance teams or sustainability executives. The key value proposition lies in explainable AI integration, allowing users to not only view predictive outputs but also understand the drivers behind elevated methane risks. Features such as facility density, regional classifications, and interaction-based climate variables can be embedded into visual interfaces, enabling early warning and scenario testing.

However, real-world deployment would require retraining with updated data, robust anomaly detection layers, and seamless integration with IoT or emissions sensor feeds. Moreover, as highlighted in the literature on data-driven environmental policy, the risk of over-reliance on opaque AI systems must be countered through transparent audit trails and public-sector standards for interpretability and fairness.

5.3 Methodological and Conceptual Insights

This project contributes to the evolving field of environmental data science by combining diverse climate, spatial, and industrial datasets to forecast methane emissions at a facility level. Through transparent feature design and interpretable modelling, it provides a framework for forecasting emissions in a way that is both rigorous and operationally applicable.

5.3.1 Data Fusion and Feature Design

A key strength of this project lies in its multi-dimensional dataset. By combining gridded NetCDF climate data with point-source emissions and contextual facility information, the modelling approach captures complex local conditions that shape methane output. Seasonal climate features were prioritised due to the known influence of temperature, wind, and atmospheric pressure on CH_4 behaviour, while spatial metrics, such as facility density and regional classification provided geographic grounding.

Interaction terms between climate variables were introduced to account for potential non-linear dynamics (e.g., wind moderating the effect of pressure). This emphasis on capturing both standalone effects and latent interactions offers a replicable structure for future emissions models that require flexibility across geographies or pollutant types.

5.3.2 Interpretability and Predictive Transparency

The application of SHAP values to the XGBoost framework ensured interpretability across both global and local dimensions. Rather than treating explainability as a post-modelling step, it was embedded into the pipeline to guide feature selection, prioritise strategic models, and evaluate the real-world drivers of emissions. This approach is particularly important in domains where black-box models can obscure key risk indicators or policy levers.

SHAP revealed that emissions were most heavily influenced by sectoral classification, spatial clustering, and selected climate indicators (e.g., winter temperature, mean wind speed). These findings validate the importance of combining domain knowledge with data-driven techniques when developing models

for regulatory or sustainability applications. Similar approaches have been used to quantify morphological or meteorological drivers of environmental risk using SHAP and XGBoost, such as urban flooding (Wang et al., 2023) and ozone pollution (Liu et al., n.d.).

5.3.3 Positioning within the Existing Literature

In contrast to many national-scale or industry-aggregated CH_4 forecasting studies, this project operates at the facility level, allowing for highly granular analysis. The integration of spatially explicit predictors, seasonal variables, and explainable AI adds to a growing body of work advocating for localised, interpretable models in environmental analytics.

The methodology also offers a transferable template. While this study focuses on methane, the same structure data fusion, seasonal aggregation, SHAP-based interpretation could be extended to CO_2, NOx, or $PM_{2.5}$ forecasting efforts, particularly in policy-sensitive or compliance-heavy contexts.

5.4 Limitations and Critical Evaluation

5.4.1 Methodological Constraints

While the model delivered strong predictive performance ($R^2 = 0.73$ on the test set), several methodological limitations must be acknowledged. Most notably, non-exploration of alternative modelling paradigms, such as deep neural networks (e.g., LSTM) represents a missed opportunity for capturing temporal autocorrelation and non-linear seasonal shifts in emissions. Prior work by Luo et al. (2024) demonstrated the effectiveness of multivariate LSTM models for methane prediction tasks, revealing enhanced short-term trend capture in volatile emission scenarios.

Moreover, the decision to prioritise explainability (via XGBoost and SHAP) constrained the model space by favouring tree-based ensembles over less interpretable alternatives. Hybrid methods such as model stacking or neural-symbolic systems might offer improved performance while retaining some interpretability (Bassil et al., 2025).

5.4.2 Data Quality and Structural Challenges

The spatial resolution mismatch between 1 km² NetCDF climate grids and exact facility coordinates introduce uncertainty, particularly in dense urban or industrial clusters. While spatial joins were performed, edge effects and buffer assignment errors may have diluted localised climate-emission signals. Additionally, the annual aggregation of emissions obscures short-term spikes, potentially underestimating the influence of transient climatic events.

The temporal granularity mismatch, monthly climate features paired with annual emissions raises concerns about the true temporal alignment between predictors and targets. This misalignment limits the ability to capture intra-annual variability in emission behaviour.

5.4.3 Statistical and Explainability Considerations

SHAP values were instrumental in model interpretability; however, their limitations must be acknowledged. When features are highly correlated, such as sector encodings and spatial indicators, SHAP may distribute importance unevenly, or fail to capture interaction dominance. Huang and Marques-Silva (2024) caution that Shapley-based explanations can mislead in contexts with multicollinearity or interdependent features, a condition likely present in this study's sectoral variables.

Further, while the model's residuals were unbiased on average, variance heteroskedasticity persisted for certain industrial categories, particularly those with extreme emissions (e.g., Major Power Producers). This suggests the need for more tailored residual diagnostics or stratified modelling approaches in future iterations.

5.5 Alternative Explanations and Robustness Considerations

While the XGBoost model using strategic features yielded robust predictive performance ($R^2 = 0.725$ on the test set), alternative modelling approaches and feature representations were considered and merit critical reflection.

5.5.1 Model Selection and Algorithm Comparison

Random Forest was evaluated as a benchmark model and performed well ($R^2 \approx 0.70$), but XGBoost consistently outperformed it across all metrics, especially in

terms of generalisation. However, other advanced approaches — including deep learning architectures such as LSTMs or hybrid temporal-convolutional models — were not explored due to limited temporal resolution (annual emissions) and the project's interpretability constraints. It remains plausible that with denser time-series emissions data, such methods could outperform gradient-boosted trees, particularly in capturing latent temporal dynamics.

Moreover, the exclusive reliance on decision tree ensembles could introduce bias toward feature importance measures such as gain or cover, even when using SHAP for post-hoc explanation. Ensemble diversity or alternative algorithms (e.g., LightGBM, CatBoost, or GAMs) may yield complementary insights or better handle categorical-spatial feature interactions.

5.5.2 Feature Engineering Alternatives

The feature engineering strategy relied on domain-guided aggregation (seasonal climate metrics) and interaction term construction. While effective, this approach may overlook latent patterns that could be automatically captured via deep feature synthesis or representation learning. Additionally, while interaction terms (e.g., pressure × wind) improved model accuracy, their selection was guided by intuition rather than automated discovery techniques (e.g., polynomial feature expansion or mutual information).

Spatial clustering indicators were limited to circular buffers (10km, 25km) , alternative geostatistical or kernel density estimations may yield more nuanced spatial influences. Similarly, sectoral features were treated equally through one-hot encoding, yet hierarchical encoding (e.g., ISIC levels) might preserve industry similarities more effectively.

5.6 Future Research Directions and Technical Enhancements

The current project establishes a strong foundation for CH_4 emissions forecasting at facility level, but there are several technical and conceptual avenues for future enhancement. These can be grouped into short-term improvements and long-term research extensions.

5.6.1 Short-Term Technical Improvements

A key limitation of the current modelling approach is its reliance on static, annualised features. Future iterations could incorporate higher temporal resolution climate and operational data (e.g., monthly or daily facility activity levels), allowing the models to detect more dynamic fluctuations in methane output. For example, introducing time-aware machine learning models such as LSTM or Transformer-based architectures could improve temporal sensitivity and anomaly detection (Pölz et al., 2024; Shi et al., 2024).

Further refinements could include enhanced feature engineering through automated techniques such as genetic algorithms or automated feature synthesis pipelines. These would complement SHAP-based interpretability tools and reduce manual bias in the feature selection process.

5.6.2 Long-Term Research Opportunities

One promising avenue is the incorporation of satellite-based remote sensing data to augment ground-level emissions reporting. Advances in atmospheric trace gas retrievals offer the potential for continuous, non-intrusive methane monitoring at regional scales. Combining these spatially expansive datasets with point-source level modelling could support multi-scale validation frameworks.

There is also scope to investigate hybrid modelling approaches that combine deep learning models for temporal patterns with tree-based explainable models for interpretability. Recent work suggests that transformer–LSTM hybrids outperform traditional recurrent models in environmental forecasting tasks due to their ability to learn long-range dependencies efficiently (Kow et al., 2024).

5.6.3 Broader Environmental Applications

The pipeline developed here can be adapted to other greenhouse gases such as CO_2 or N_2O, provided equivalent emissions inventories and environmental covariates are available. Applying the methodology across different sectors (e.g., agriculture, transportation) and geographies would test the generalisability of the framework and its potential as a broader emissions intelligence tool. Such cross-

context replication is crucial for building a robust environmental data science toolkit that supports both policy and operational sustainability goals.

6. Conclusion and Recommendations

6.1 Summary of Key Findings

This dissertation developed and evaluated a machine learning-based framework to predict facility-level methane (CH_4) emissions in the UK by fusing spatial, sectoral, and climate data from 2018 to 2022. In response to the increasing urgency of addressing short-lived climate pollutants, and GreentecAI's goal of operationalizing scalable emissions forecasting, this study implemented a predictive pipeline that integrated point-source emissions records with high-resolution gridded meteorological variables and spatial context indicators.

The XGBoost model, trained on a strategically engineered subset of features, emerged as the most effective algorithm in terms of predictive accuracy and generalisability. The final tuned model achieved a test R^2 of 0.72 and a MAE of 0.27, outperforming baseline models while maintaining interpretability through SHAP (SHapley Additive exPlanations) analysis. Strategic feature selection played a pivotal role in reducing overfitting, with key predictors including winter temperature, mean wind speed, regional classifications, facility density within 25 km, and sectoral identifiers.

Exploratory Data Analysis (EDA) revealed substantial heterogeneity in emissions patterns across regions, sectors, and urban–rural designations. The Oil and Gas, Power Generation, Waste, and Chemical sectors were consistently the highest emitters, while climate overlays highlighted how emissions hotspots tended to co-locate with regions exhibiting elevated temperatures and wind speeds. Correlation analysis further demonstrated that climatic and spatial variables, particularly facility clustering and winter conditions, were non-trivial contributors to emission variability.

SHAP values reinforced these findings, assigning high importance to spatial proximity metrics and sectoral identifiers. The interpretability of these results provides actionable insight for stakeholders seeking to identify super-emitters, optimise monitoring, and target policy interventions.

6.2 Contributions to Environmental Analytics

This project advances the field of data-driven environmental monitoring in several keyways:

- **Fine-grained emissions forecasting**: By integrating 1 km² gridded climate data with facility-level emissions, the model enables localised prediction of methane output, surpassing the coarse granularity of most national inventories or satellite-based estimates.

- **Multi-source data fusion**: The methodological pipeline addressed the challenges of aligning temporally and spatially heterogeneous datasets. Unlike many prior studies that rely on either top-down (remote sensing) or bottom-up (inventory) data alone, this study demonstrated the value of combining both emissions data and contextual environmental signals to improve accuracy and relevance.

- **Explainable ML for policy**: The incorporation of SHAP analysis makes the model interpretable and thus suitable for practical use in regulatory, compliance, and planning contexts. Insights from feature importance rankings can inform sector-specific abatement strategies and help target inspections or retrofit initiatives.

- **Transferability and scalability**: While this study focused on methane emissions from UK facilities, the pipeline design is generalisable. With appropriate datasets, this framework can be applied to other greenhouse gases, geographies, or even mobile sources (e.g., transport networks), provided spatial and environmental inputs are available.

6.3 Limitations

Despite its contributions, the project is subject to several limitations:

1. **Temporal resolution**: Emissions data were reported annually, while climate variables were averaged at monthly or seasonal levels. This temporal mismatch may obscure short-term weather-driven emission spikes or anomalous facility events (e.g., leaks, maintenance activities). Real-time sensor integration would enhance predictive responsiveness.

2. **Static point-source assumption**: The model treats facility locations and operational status as fixed across the five-year period, which may not reflect temporary closures, upgrades, or policy-induced changes in production. This simplification, while necessary for alignment, could affect model precision for newer or transient emitters.

3. **Lack of process-level granularity**: Facility-level emissions are aggregated totals, without differentiation between methane-generating processes (e.g., flaring, venting, biological decomposition). Including process-level data could improve attribution and yield more targeted recommendations.

4. **Exclusion of agriculture**: Agricultural methane sources, especially livestock and anaerobic digestion, were excluded due to incompatible spatial resolution. This limits the comprehensiveness of the national methane emissions picture, although it aligns with GreentecAI's current focus on industrial point sources.

5. **Potential spatial bias**: While climate and emissions data were harmonised geographically, measurement density may still vary by region, with urban areas likely overrepresented relative to rural or remote zones.

6.4 Practical Recommendations

Based on the findings and limitations, the following recommendations are made for GreentecAI and similar organisations aiming to deploy CH_4 forecasting tools:

1. **Deploy SHAP-enhanced dashboards**: Integrate SHAP values into operational interfaces to allow end-users to understand not only what the model predicts, but why. This enhances transparency, supports regulatory justification, and builds trust in AI-powered decisions.

2. **Incorporate real-time sensor data**: Extending the pipeline to include Internet of Things (IoT) methane sensors would enable real-time emissions forecasting, especially valuable in industrial sites with frequent operational changes or leak risks.

3. **Use facility density indicators in inspection prioritisation**: Areas with low facility count but high emissions (e.g., remote oil fields) warrant targeted oversight, while regions with high density but moderate emissions could benefit from networked monitoring strategies.

4. **Embed the model in early warning systems**: With appropriate temporal extensions, the predictive model could serve as the core of an early warning system that alerts operators to climate-emission conditions likely to result in exceedances or hotspots.

5. **Align model insights with net-zero roadmaps**: Use the feature importance insights to design sector-specific decarbonisation pathways, particularly for oil and gas operators, where policy levers (e.g., leak detection mandates) can

have high impact.

6. **Invest in higher-resolution temporal and process-level data**: Future iterations of the model should aim to integrate monthly or daily emissions readings and break down methane contributions by activity type, enabling more nuanced forecasting and policy alignment.

6.5 Directions for Future Research

To build upon this work and address the identified limitations, several research directions are proposed:

- **Multitarget learning**: Extend the framework to predict other greenhouse gases (e.g., CO_2, N_2O) alongside methane. This would support integrated emissions forecasting and better inform sustainability planning.

- **Spatiotemporal deep learning**: Explore LSTM or transformer-based architectures capable of capturing temporal dependencies and spatial autocorrelation. This may improve performance on non-linear and lagged climate-emissions relationships.

- **Sector-specific models**: Train models customised to key emission-heavy industries. Tailored features and tuning strategies can better capture intra-sector heterogeneity.

- **Synthetic data augmentation**: Use generative techniques or scenario simulations to expand training data under underrepresented environmental conditions, which may improve generalisation in atypical future climates.

- **Comparative policy evaluation**: Link model outputs to policy databases (e.g., regulatory enforcement actions) to assess whether predictions align with known violations or interventions, offering validation and insights for decision support.

6.6 Final Reflections

This dissertation demonstrated the feasibility and value of combining machine learning with environmental and spatial data to improve methane emissions prediction. By building a robust, interpretable forecasting pipeline and empirically validating its performance across multiple model variants, the project contributes both methodologically and operationally to the field of environmental analytics.

As the UK and global actors move toward net-zero targets, tools that enhance visibility into short-lived climate pollutants will be increasingly vital. This study provides a scalable blueprint for such systems, capable of informing local interventions, national inventories, and cross-sectoral planning. While challenges remain in data alignment, model generalisation, and interpretability, this work lays a strong foundation for real-world deployment and further research innovation.

7. GenAI Use and Critical Evaluation

7.1 Overview of GenAI Integration

Generative AI tools were strategically and transparently integrated throughout this dissertation, supporting literature review, code development, debugging, and academic writing. Tools included **ChatGPT (OpenAI)**, **Claude (Anthropic)**, **Perplexity.ai**, and **Cursor IDE**. These were used iteratively to accelerate technical tasks and enhance academic structuring , but all final judgments, selections, and interpretations were critically evaluated and authored independently.

7.2 Literature Review Support via Perplexity.ai

To streamline literature discovery, I used Perplexity.ai, a GenAI-powered academic search engine, to identify recent peer-reviewed research on methane emissions forecasting, SHAP interpretation, and environmental machine learning. It helped highlight relevant authors and journals (e.g., *Scientific Reports*, *Environmental Management*), which were then verified manually via Google Scholar and publisher databases.

Perplexity often provided quick summaries and suggested DOIs, but I cross-checked each source to avoid fabricated citations, a known issue in GenAI tools. Its main value lay in speeding up literature mapping and helping me formulate thematic subsections (e.g., facility-level forecasting, AI explainability in climate).

7.3 Programming and Debugging with Cursor IDE, ChatGPT, and Claude

For coding tasks including data cleaning, spatial joins, feature engineering, SHAP integration, and model evaluation, I used Cursor IDE, an AI-native development environment that integrates real-time ChatGPT assistance into the coding workflow. Cursor was used to:

- Quickly scaffold boilerplate functions
- Refactor redundant code blocks

- Suggest debugging fixes (e.g., for SHAP plotting errors)

In parallel, I used ChatGPT and Claude to generate or compare code snippets, understand unexpected outputs, and fine-tune pipelines. Their suggestions were critically evaluated and modified in context using domain knowledge and experimental testing.

No code was directly copied without adaptation. All final implementations were tested and tuned within Jupyter Notebooks with logic, structure, and evaluation decisions made by me.

7.4 Dissertation Structuring and Writing Support

I used ChatGPT and Claude as structural collaborators , particularly during:

- Outlining dissertation chapters
- Refining the flow of SHAP analysis and methodological justifications
- Generating alternative phrasings or transitions to improve clarity

I prompted both models to critique draft sections and often compared their responses to decide which suggestions to adopt. However, AI was only used to improve expression, not to generate original analysis. All critical insights (e.g., sectoral SHAP interpretations, model comparisons) were based on my interpretation of my results.

7.5 Ethical and Academic Integrity

I adhered strictly to UCL’s academic integrity and GenAI guidelines by:

- Critically reviewing all GenAI-generated content
- Verifying factual claims and citations
- Clearly distinguishing AI-assisted text from my original work

No sensitive data was shared with AI systems, and no outputs were submitted without human validation and rewriting.

7.6 Reflection and Limitations

Overall, GenAI significantly enhanced the efficiency and quality of the research process, particularly in early-stage ideation, exploratory coding, and academic structuring.

- **Perplexity.ai** occasionally surfaced non-existent citations
- **Cursor IDE** sometimes misunderstood context, offering inefficient solutions
- **ChatGPT/Claude** occasionally produced generic or verbose phrasing

These limitations underscored the importance of treating GenAI as a support tool not a substitute for domain knowledge, empirical insight, or ethical reasoning.in future research, I would continue to use GenAI for exploration and rapid prototyping, but always within a critical, human-in-the-loop framework.

8.References

Akiba, T., Sano, S., Yanase, T., Ohta, T. and Koyama, M. (2019) 'Optuna: A Next-generation Hyperparameter Optimization Framework', Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. doi: https://doi.org/10.1145/3292500.3330701.

Alam, G.M.I. et al. (2025) 'Deep learning model based prediction of vehicle CO_2 emissions with eXplainable AI integration', Scientific Reports, 15(1). doi: https://doi.org/10.1038/s41598-025-87233-y.

Alnuaimi, H.S. et al. (2025) 'Machine Learning Applications for Carbon Emission Estimation', Resources, Conservation & Recycling Advances, p.200263. doi: https://doi.org/10.1016/j.rcradv.2025.200263.

Balać, N. et al. (2025) 'Implementation of XGBoost Models for Predicting CO_2 Emission and Specific Tractor Fuel Consumption', Agriculture, 15(11), p.1209. doi: https://doi.org/10.3390/agriculture15111209.

Bassil, J., Noura, H.N., Salman, O., Chahine, K. and Guizani, M. (2025) 'Efficient combination of deep learning and tree-based classification models for solar panel dust detection', Intelligent Systems with Applications, 26, p.200509. doi: https://doi.org/10.1016/j.iswa.2025.200509.

Bhardwaj, P., Kumar, R., Mitchell, D.A., Randles, C.A., Downey, N., Blewitt, D. and Kosovic, B. (2022) 'Evaluating the detectability of methane point sources from satellite observing systems using microscale modeling', Scientific Reports, 12(1). doi: https://doi.org/10.1038/s41598-022-20567-z.

Bista, P. et al. (2025) 'Evaluating Machine Learning Models for Greenhouse Gas Emissions Prediction in Diversified Semi-Arid Cropping Systems', Soil Science Society of America Journal, 89(2). doi: https://doi.org/10.1002/saj2.70057.

Chai, T. and Draxler, R.R. (2014) 'Root mean square error (RMSE) or mean absolute error (MAE)?', Geoscientific Model Development, 7(3), pp.1247–1250. doi: https://doi.org/10.5194/gmd-7-1247-2014.

Chang, H.-T., Chern, Y.-R., Asri, A.K., Liu, W.-Y., Hsu, C.-Y., Hsiao, T.-C., Chi, K.H., Lung, S.-C.C. and Wu, C.-D. (2025) 'Innovating Taiwan's greenhouse gas estimation: A case study of atmospheric methane using GeoAI-based ensemble mixed spatial prediction model', Journal of Environmental Management, 380, p.125110. doi: https://doi.org/10.1016/j.jenvman.2025.125110.

Cionni, I., Lledó, L., Torralba, V. and Dell'Aquila, A. (2022) 'Seasonal predictions of energy-relevant climate variables through Euro-Atlantic Teleconnections', Climate Services, 26, p.100294. doi: https://doi.org/10.1016/j.cliser.2022.100294.

Cutler, J., Li, B., Alhnaity, B., Partridge, T., Thompson, M. and Meng, Q. (2025) 'AI for sustainable land management and greenhouse gas emission forecasting: advancing climate action', IEEE CACML, pp.1–8. doi: https://doi.org/10.1109/CACML64929.2025.11010973.

Esiri, B. and Ekemezie, P. (2024) 'Standardizing methane emission monitoring: A global policy perspective for the oil and gas industry', Engineering Science & Technology Journal, 5(6), pp.2027–2038. doi: https://doi.org/10.51594/estj/v5i6.1220.

GeeksforGeeks (2019) One Hot Encoding in Machine Learning. Available at: https://www.geeksforgeeks.org/machine-learning/ml-one-hot-encoding/ (Accessed: 10 July 2025).

Hasan, R. et al. (2025) 'AI-Driven Greenhouse Gas Monitoring: Enhancing Accuracy, Efficiency, and Real-Time Emissions Tracking', AIMS Environmental Science, 12(3), pp.495–525. doi: https://doi.org/10.3934/environsci.2025023.

Huang, X. and Marques-Silva, J. (2024) 'On the failings of Shapley values for explainability', International Journal of Approximate Reasoning, 171, p.109112. doi: https://doi.org/10.1016/j.ijar.2023.109112.

IPCC (2021) Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Available at:

https://www.ipcc.ch/report/ar6/wg1/downloads/report/IPCC_AR6_WGI_SPM.pdf (Accessed: 7 July 2025).

Jeong, S., Hamilton, S.D., Johnson, M.S., Wu, D., Turner, A.J. and Fischer, M.L. (2025) 'Applying Gaussian Process Machine Learning and modern probabilistic programming to satellite data to infer CO_2 emissions', Environmental Science & Technology. doi: https://doi.org/10.1021/acs.est.4c09395.

Kow, P.-Y., Liou, J.-Y., Yang, M.-T., Lee, M.-H., Chang, L.-C. and Chang, F.-J. (2024) 'Advancing climate-resilient flood mitigation: Utilizing transformer-LSTM for water level forecasting at pumping stations', Science of The Total Environment, 927, p.172246. doi: https://doi.org/10.1016/j.scitotenv.2024.172246.

Lei, T., Chen, X., Ma, S., Guan, D. and Jing, L. (2025) 'A global inventory of methane emissions from abandoned oil and gas wells and possible mitigation pathways', National Science Review, 12(7). doi: https://doi.org/10.1093/nsr/nwaf184.

Li, Z., Wang, Y., Liu, J. and Xian, J. (2025) 'Using machine learning to unravel chemical and meteorological effects on ground-level ozone: Insights for ozone-climate control strategies', Environment International, 201, p.109567. doi: https://doi.org/10.1016/j.envint.2025.109567.

Liu, Z., Lu, Z., Zhu, W., Yuan, J., Cao, Z., Cao, T., Liu, S., Xu, Y. and Zhang, X. (2025) 'Comparison of machine learning methods for predicting ground-level ozone pollution in Beijing', Frontiers in Environmental Science, 13, p.1561794. doi: https://doi.org/10.3389/fenvs.2025.1561794.

Luo, R., Wang, J. and Gates, I. (2024) 'Forecasting Methane Data Using Multivariate Long Short-Term Memory Neural Networks', Environmental Modeling & Assessment, 29(3), pp.441–454. doi: https://doi.org/10.1007/s10666-024-09957-x.

Mathew, N., Somanathan, A., Tirpude, A. and Arfin, T. (2024) 'The impact of short-lived climate pollutants on the human health', Environmental Pollution and Management, 1, pp.1–14. doi: https://doi.org/10.1016/j.epm.2024.04.001.

Pölz, A., Blaschke, A.P., Komma, J., Farnleitner, A.H. and Derx, J. (2024) 'Transformer Versus LSTM: A Comparison of Deep Learning Models for Karst Spring Discharge Forecasting', Water Resources Research, 60(4). doi: https://doi.org/10.1029/2022wr032602.

Saunois, M., Stavert, A.R., Poulter, B., Bousquet, P., Canadell, J.G., Jackson, R.B., Raymond, P.A., Dlugokencky, E.J., Houweling, S., Patra, P.K., Ciais, P., Arora, V.K., Bastviken, D., Bergamaschi, P., Blake, D.R., Brailsford, G., Bruhwiler, L., Carlson, K.M., Carroll, M. and Castaldi, S. (2020) 'The Global Methane Budget 2000–2017', Earth System Science Data, 12(3), pp.1561–1623. doi: https://doi.org/10.5194/essd-12-1561-2020.

Shaibu, S. (2024) Normalization vs. Standardization: How to Know the Difference. DataCamp. Available at: https://www.datacamp.com/tutorial/normalization-vs-standardization (Accessed: 10 July 2025).

Sherwin, E.D., Rutherford, J.S., Chen, Y., Aminfard, S., Kort, E.A., Jackson, R.B. and Brandt, A.R. (2023) 'Single-blind validation of space-based point-source detection and quantification of onshore methane emissions', Scientific Reports, 13(1), p.3836. doi: https://doi.org/10.1038/s41598-023-30761-2.

Shi, J., Wang, S., Qu, P. and Shao, J. (2024) 'Time series prediction model using LSTM-Transformer neural network for mine water inflow', Scientific Reports, 14(1). doi: https://doi.org/10.1038/s41598-024-69418-z.

Si, M. et al. (2024) 'Long-Term Evaluation of Machine Learning Based Methods for Air Emission Monitoring', Environmental Management, 75(3), pp.680–693. doi: https://doi.org/10.1007/s00267-024-02057-2.

Smeeton, G. (2022) Net zero: why is it necessary? Energy & Climate Intelligence Unit. Available at: https://eciu.net/analysis/briefings/net-zero/net-zero-why (Accessed: 7 July 2025).

Stecher, L., Winterstein, F., Jöckel, P., Ponater, M., Mertens, M. and Dameris, M. (2025) 'Chemistry–climate feedback of atmospheric methane in a methane-emission-

flux-driven chemistry–climate model', Atmospheric Chemistry and Physics, 25(10), pp.5133–5158. doi: https://doi.org/10.5194/acp-25-5133-2025.

Tolou Shadbahr, S., Roberts, M., Stanczuk, J., Gilbey, J., Teare, P., Dittmer, S., Thorpe, M., Ramon Viñas Torné, Sala, E., Lió, P., Patel, M., Preller, J., Selby, I., Breger, A., Weir-McCall, J.R., Gkrania-Klotsas, E., Korhonen, A., Jefferson, E., Langs, G. and Yang, G. (2023) 'The impact of imputation quality on machine learning classifiers for datasets with missing values', Communications Medicine, 3(1). doi: https://doi.org/10.1038/s43856-023-00356-z.

Uppalapati, S., Paramasivam, P., Kilari, N., Chohan, J.S., Kanti, P.K., Vemanaboina, H., Dabelo, L.H. and Gupta, R. (2025) 'Precision biochar yield forecasting employing random forest and XGBoost with Taylor diagram visualization', Scientific Reports, 15(1). doi: https://doi.org/10.1038/s41598-025-91450-w.

Vollrath, C., Xing, Z., Hugenholtz, C., Barchyn, T. and Winter, J. (2024) 'A review of methane emissions source types, characteristics, rates, and mitigation across U.S. and Canadian cities', ChemRxiv. doi: https://doi.org/10.26434/chemrxiv-2024-cqzpr-v2.

Wang, M., Li, Y., Yuan, H., Zhou, S., Wang, Y., Muhammad, R., Ikram, A. and Li, J. (2023) 'An XGBoost–SHAP approach to quantifying morphological impact on urban flooding susceptibility', Ecological Indicators, 156, p.111137. doi: https://doi.org/10.1016/j.ecolind.2023.111137.

West, R.M. (2021) 'Best practice in statistics: The use of log transformation', Annals of Clinical Biochemistry: International Journal of Laboratory Medicine, 59(3), pp.162–165. doi: https://doi.org/10.1177/00045632211050531.

Xu, Y., Yazdinejad, A., Wang, H. and Kong, J.D. (2025) 'Methane monitoring: A systematic review of multi-source data integration challenges and solutions', SSRN. doi: https://doi.org/10.2139/ssrn.5218738.

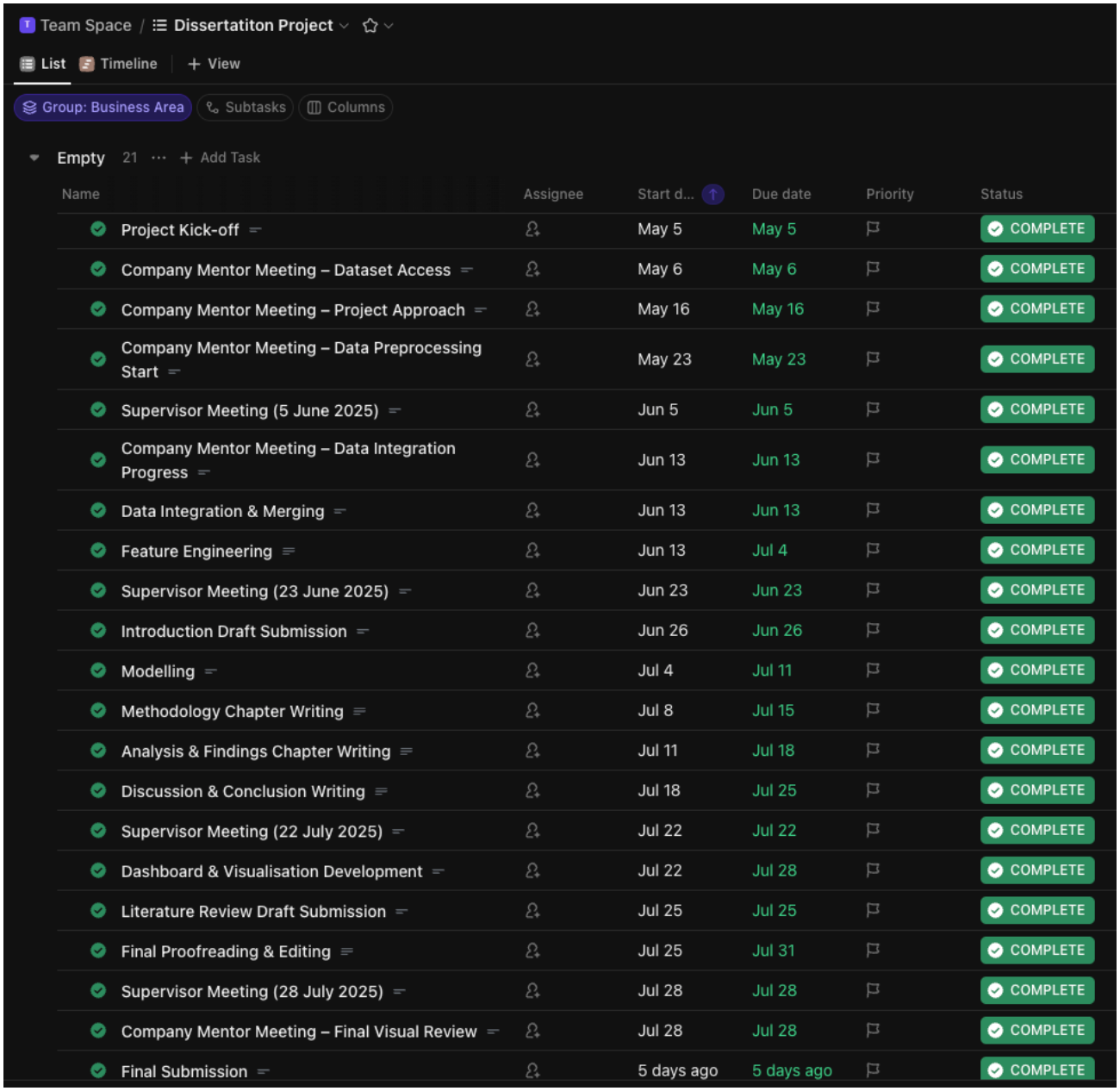
Yazdinejad, A., Wang, H. and Kong, J.D. (2025) 'Advanced AI-driven methane emission detection, quantification, and localization in Canada: A hybrid multi-source fusion framework', SSRN. doi: https://doi.org/10.2139/ssrn.5223515.

Zhang, S., Ma, J., Zhang, X. and Guo, C. (2023) 'Atmospheric remote sensing for anthropogenic methane emissions: Applications and research opportunities', Science of The Total Environment, 893, p.164701. doi: https://doi.org/10.1016/j.scitotenv.2023.164701.

9. Appendix

Appendix A: Project Management

https://app.clickup.com/90151460220/v/li/901513927536



Team Space / Dissertatiton Project

List | Timeline | + View

Group: Business Area | Subtasks | Columns

Empty 21 ··· + Add Task

Name	Assignee	Start d...	Due date	Priority	Status
Project Kick-off		May 5	May 5		COMPLETE
Company Mentor Meeting – Dataset Access		May 6	May 6		COMPLETE
Company Mentor Meeting – Project Approach		May 16	May 16		COMPLETE
Company Mentor Meeting – Data Preprocessing Start		May 23	May 23		COMPLETE
Supervisor Meeting (5 June 2025)		Jun 5	Jun 5		COMPLETE
Company Mentor Meeting – Data Integration Progress		Jun 13	Jun 13		COMPLETE
Data Integration & Merging		Jun 13	Jun 13		COMPLETE
Feature Engineering		Jun 13	Jul 4		COMPLETE
Supervisor Meeting (23 June 2025)		Jun 23	Jun 23		COMPLETE
Introduction Draft Submission		Jun 26	Jun 26		COMPLETE
Modelling		Jul 4	Jul 11		COMPLETE
Methodology Chapter Writing		Jul 8	Jul 15		COMPLETE
Analysis & Findings Chapter Writing		Jul 11	Jul 18		COMPLETE
Discussion & Conclusion Writing		Jul 18	Jul 25		COMPLETE
Supervisor Meeting (22 July 2025)		Jul 22	Jul 22		COMPLETE
Dashboard & Visualisation Development		Jul 22	Jul 28		COMPLETE
Literature Review Draft Submission		Jul 25	Jul 25		COMPLETE
Final Proofreading & Editing		Jul 25	Jul 31		COMPLETE
Supervisor Meeting (28 July 2025)		Jul 28	Jul 28		COMPLETE
Company Mentor Meeting – Final Visual Review		Jul 28	Jul 28		COMPLETE
Final Submission		5 days ago	5 days ago		COMPLETE

Appendix B: Predictive Modelling Notebook

- *The complete codebase for data preprocessing, feature engineering, modelling, and evaluation is available via Google Colab. The notebook can be accessed using the link below:*

https://colab.research.google.com/drive/1DeAeOeH59S3Nao0_LvaP8sije02pvDHW?usp=sharing