



AI-Based Nowcasting of National CO₂ Emissions
Using Satellite-Derived Indicators:
A Hybrid Framework for GCC Countries

Prepared By
Mohammad Wajeeh Bader

Student ID: 251000462

Submitted in fulfillment of the requirements for the degree Master
of Science in Artificial Intelligence

Supervised by
Dr. Amjad Qtaish

College of Informatics, Midocean University, Comoros

November 2025

Acknowledgments

I could not have completed this work without God's grace and the support of many people who helped me along the way.

My deepest thanks go to my academic supervisor. When I started this project, I felt lost, overwhelmed by the datasets and not sure where to begin. My supervisor didn't just help me process data. He taught me how to think as a researcher: how to structure ideas, synthesize literature, and apply rigorous methods. He pushed me when I needed it and never let me quit. Thank you for your patience and for helping me find my footing.

I also want to recognize Google Earth Engine, which transformed my research. Without it, I'd probably still be downloading MODIS tiles manually. Thanks to Google, NASA, ESA, and NOAA for making 24 years of satellite data accessible and free. Similarly, Our World in Data and the Global Carbon Project saved me weeks of work by providing clean emissions data ready to use.

To my colleagues at my workplace, sorry for disappearing so often, and thanks for putting up with my endless talk about the latest algorithm updates.

Finally, my family. They don't really understand what machine learning is or why I care about carbon emissions, but they supported me anyway. My father kept asking when I'd be done, which somehow kept me going. My mother woke me up every morning, prayed for me, and made sure I had everything I needed. This thesis happened because they believed in me.

Abstract

National greenhouse gas inventories are generally published with a one- to two-year lag following the actual emission events, causing governments to base policy choices on information that is no longer current. This study investigates whether satellite-based observations can provide timely estimates of national CO₂ emissions for Gulf Cooperation Council (GCC) countries, potentially narrowing this temporal gap.

For Saudi Arabia, the United Arab Emirates, Bahrain, Kuwait, Oman, and Qatar, I compiled a balanced panel dataset spanning 2000–2023 (144 country-year observations). Production-based CO₂ data and macroeconomic indicators come from the Our World in Data (OWID) database. Satellite-derived metrics are extracted from five Google Earth Engine products: MODIS land surface temperature (LST), MODIS vegetation index (NDVI), DMSP/VIIRS night-time lights, MERRA-2 aerosol optical depth (AOD), and Sentinel-5P tropospheric NO₂. All variables are aggregated to annual country-year format for cross-source comparability.

Three XGBoost regression architectures were constructed and assessed through temporal validation. Training data encompasses 2000–2018, with 2019–2023 reserved for testing. Model A uses only macroeconomic and demographic variables (GDP, population, energy consumption) as an inventory-based reference. Model B merges remote sensing indicators with population density information. Model C operates solely on satellite-based measurements. Throughout the test phase, Model A attains an R² of approximately 0.96, demonstrating that conventional economic approaches perform well when recent statistics are available. The hybrid and satellite-only models reach test R² values of approximately 0.69 and 0.67, respectively, and still capture most of the year-to-year variations in national CO₂ emissions.

Analysis reveals that seasonal land surface temperatures (summer and winter), nocturnal luminosity, and demographic density constitute the dominant predictive features. These findings correspond to the primary emission sources across GCC nations, where elevated temperatures, concentrated urbanization, and intensive power usage substantially shape national carbon dioxide outputs. Overall, satellite-based models cannot fully substitute for official inventories; nevertheless, they provide rapid, physically grounded nowcasts that complement conventional methods and reduce delays in emissions tracking.

Keywords: CO₂ emissions, machine learning, XGBoost, satellite remote sensing, Google Earth Engine, land surface temperature, night-time lights, MODIS, VIIRS, GCC, climate monitoring, nowcasting

Table of Contents

Acknowledgments	i
Abstract.....	ii
Table of Contents.....	iii
List of Abbreviations	v
List of Figures.....	vi
List of Tables	vii
Chapter 1 Introduction.....	8
1.1 Introduction.....	8
1.2 Problem Definition	9
1.3 Thesis Objectives.....	10
1.4 Thesis Contribution.....	10
1.5 Thesis Organization	11
Chapter 2 Overview	12
2.1 Introduction.....	12
2.2 Artificial Intelligence.....	12
2.3 Machine Learning	13
2.4 Supervised Learning in Detail	15
2.5 Deep Learning.....	16
2.6 Ensemble Methods.....	17
2.7 The XGBoost Algorithm.....	18
2.8 Greenhouse Gas Emissions and Climate Policy	18
2.9 The GCC Region: Energy Profile and Climate Commitments	19
2.10 Earth Observation for Environmental Monitoring	21
2.11 Satellite Data Sources Used in This Study.....	21
2.12 Machine Learning for Emission Estimation	23
2.13 XGBoost: Mathematical Formulation.....	23
2.14 Explainable AI for Tree-Based Models	25
Chapter 3 Related Work	26
3.1 Introduction.....	26
3.2 AI-Based Emission Forecasting.....	27
3.3 Satellite-Based Carbon Monitoring	28
3.4 Regional Studies in the GCC	29
3.5 Research Gaps and Contributions	30
Chapter 4 Proposed Approach	31
4.1 Introduction.....	31
4.2 Research Design Overview.....	31
4.3 Study Area: The GCC Region	34
4.4 Data Sources and Collection.....	35
4.4.1 Emission Data	35

4.4.2 Economic and Demographic Data	35
4.4.3 Satellite Data.....	36
4.5 Satellite Data Extraction Pipeline	36
4.5.1 Google Earth Engine Image Collections.....	36
4.5.2 Extraction Workflow	37
4.5.3 Handling the DMSP–VIIRS Transition	38
4.5.4 Missing Data Handling	39
4.6 Feature Engineering and Preprocessing.....	39
4.6.1 Seasonal Disaggregation of Temperature	39
4.6.2 Population Density Transformation.....	40
4.6.3 Final Feature Set	40
4.6.4 Correlation Analysis	41
4.7 Experimental Design: Three-Model Comparison	42
4.8 Model Training and Validation.....	43
4.8.1 Temporal Train–Test Split.....	43
4.8.2 XGBoost Configuration.....	43
4.9 Evaluation Metrics	44
Chapter 5 Experimental Results and Discussion	45
5.1 Introduction.....	45
5.2 Data Quality Assessment	45
5.3 Model Performance Comparison	47
5.4 Feature Importance Analysis	50
5.5 Error Analysis and Residual Diagnostics.....	51
5.6 Country-Level Performance.....	53
5.7 Discussion of Findings.....	55
5.8 Summary of the Chapter	56
Chapter 6 Conclusion and Future Work	57
6.1 Summary of Contributions.....	57
6.2 Practical Implications	58
6.3 Scope and Future Extensions of the Proposed Framework.....	59
6.4 Concrete Future Work Directions	61
References	63
Appendices	66
Appendix A: Google Earth Engine Code Snippets.....	66
A.1 LST Extraction Example.....	66
A.2 Night-time Lights Extraction	66
Appendix B: Dataset Variables.....	67

List of Abbreviations

AI	Artificial Intelligence
ANFIS	Adaptive Neuro-Fuzzy Inference System
AOD	Aerosol Optical Depth
CNN	Convolutional Neural Network
CO₂	Carbon Dioxide
DMSP	Defense Meteorological Satellite Program
EO	Earth Observation
ESA	European Space Agency
FFNN	Feed-Forward Neural Network
GCC	Gulf Cooperation Council
GEE	Google Earth Engine
GHG	Greenhouse Gas
IPCC	Intergovernmental Panel on Climate Change
LST	Land Surface Temperature
LSTM	Long Short-Term Memory (Neural Network)
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MERRA-2	Modern-Era Retrospective Analysis for Research and Applications, Version 2
ML	Machine Learning
MODIS	Moderate Resolution Imaging Spectroradiometer
Mt	Megatons (Million Metric Tons)
NASA	National Aeronautics and Space Administration
NDC	Nationally Determined Contribution
NDVI	Normalized Difference Vegetation Index
NOAA	National Oceanic and Atmospheric Administration
NO₂	Nitrogen Dioxide
NTL	Night-time Lights
OCO-2	Orbiting Carbon Observatory-2
OLS	Operational Linescan System
OWID	Our World in Data
R²	Coefficient of Determination
RMSE	Root Mean Squared Error
TROPOMI	Tropospheric Monitoring Instrument
UAE	United Arab Emirates
UNFCCC	United Nations Framework Convention on Climate Change
VIIRS	Visible Infrared Imaging Radiometer Suite
XGBoost	Extreme Gradient Boosting

List of Figures

Figure 2.1: Artificial Intelligence and Its Subfields.....	12
Figure 2.2: Types of Machine Learning	14
Figure 2.3: Supervised Learning Workflow	15
Figure 2.4: Neural Network Architecture	16
Figure 2.5: Ensemble Methods Comparison.....	17
Figure 2.6: Map of the GCC Study Area	19
Figure 2.7: Satellite Data Sources and Processing Pipeline	21
Figure 2.8: XGBoost Sequential Tree Building Process	24
Figure 3.1: Literature Review Timeline (2016-2025).....	26
Figure 3.2: Research Gap Analysis Diagram.....	30
Figure 4.1: Research Framework Architecture	31
Figure 4.2: GCC CO2 Emissions Distribution by Country (2023).....	34
Figure 4.3: Google Earth Engine Data Extraction Workflow.....	37
Figure 4.4: Seasonal Land Surface Temperature Pattern.....	40
Figure 4.5: Feature Correlation Matrix	41
Figure 4.6: Temporal Train-Test Split Strategy	43
Figure 5.1: Three-Model Performance Comparison.....	48
Figure 5.2: Training vs Test R ² Comparison	49
Figure 5.3: Feature Importance Analysis (Model B).....	50
Figure 5.4: Actual vs Predicted Emissions (Scatter Plot).....	51
Figure 5.5: Residual Distribution Analysis.....	52
Figure 5.6: Time Series: Actual vs Predicted (Saudi Arabia).....	53
Figure 5.7: Country-Level R ² Performance Comparison.....	54
Figure 6.1: Accuracy vs Data Timeliness Trade-off.....	58
Figure 6.2: Future Research Directions	61

List of Tables

Table 2.1: Comparison of AI, ML, and DL	17
Table 2.2: GCC Countries Climate Targets and Emission Profiles.....	20
Table 2.3: Satellite Sensors and Products Used in This Study	23
Table 3.1: Summary of AI-Based Emission Forecasting Studies.....	28
Table 3.2: Summary of Satellite-Based Carbon Monitoring Studies	29
Table 4.1: Study Area Overview - Six GCC Countries	34
Table 4.2: Data Sources and Temporal Coverage	36
Table 4.3: Google Earth Engine Image Collections Used	36
Table 4.4: Satellite Feature Extraction Parameters.....	38
Table 4.5: Missing Data Summary	39
Table 4.6: Seasonal Aggregation Windows.....	39
Table 4.7: Final Feature Set for Each Model.....	40
Table 4.8: Three-Model Experimental Design	42
Table 4.9: XGBoost Hyperparameters.....	43
Table 5.1: Data Completeness Summary.....	45
Table 5.2: Descriptive Statistics of Key Variables	45
Table 5.3: Model Performance Comparison (Test Set 2019-2023).....	47
Table 5.4: Training vs Test Performance (Overfitting Check).....	48
Table 5.5: Feature Importance Rankings (Model B)	50
Table 5.6: Country-Level Model Performance	53

Chapter 1 Introduction

1.1 Introduction

Climate change is one of this century's biggest challenges. The IPCC 2023 Synthesis Report shows that keeping global warming below 1.5°C means cutting greenhouse gas emissions by approximately 43% by 2030 (relative to 2019 levels) and reaching net-zero CO₂ in the early 2050s (IPCC, 2023).

But official emission inventories show up late. Most countries publish their emission inventories with a 1–2-year lag, meaning that emissions from 2023 typically appear in official reports around late 2024 or early 2025. Policymakers work with old information, making it hard to respond to new emission patterns.

These delays go beyond paperwork. Managing emissions needs feedback. When data arrives late, governments can't tell if policies work or shift direction when patterns change. By the time official numbers confirm a problem, the window for fast action may be closed.

Satellite remote sensing could help. Earth observation systems watch surface temperature, nighttime lights in cities, how vegetation grows, and what's in the atmosphere. Satellite data is ready in days, not years. If researchers can show solid connections between satellite observations and actual carbon emissions, we could get near-real-time estimates without waiting for official reports.

This thesis tests this approach for the Gulf Cooperation Council (GCC): Saudi Arabia, UAE, Qatar, Kuwait, Oman, and Bahrain. These six countries work well as a test case. Hot, dry weather means heavy electricity use for cooling in summer. As major oil and gas producers, they have some of the world's highest per-person emissions. All six made climate promises under the Paris Agreement, giving them reasons to improve emission monitoring.

1.2 Problem Definition

This thesis confronts the lag between when emissions happen and when they appear in official records. Traditional emission accounting depends on economic statistics: energy consumption, industrial production, transportation activity. Collecting, verifying, and compiling these data takes substantial time, which creates the delays that limit real-time policy adjustments.

Recent work in machine learning and satellite remote sensing suggests alternatives. Research has shown that satellite-derived measurements can function as proxies for economic activity and energy use. Nighttime lights correlate with electricity consumption. Surface temperature reflects demand for cooling. Atmospheric nitrogen dioxide indicates combustion in vehicles and power plants.

Most existing studies, however, operate at either global or highly localized scales. Global models sacrifice the regional detail necessary for national policy. City-level or site-specific studies do not aggregate cleanly to country-level estimates. The GCC region adds further complications: extreme heat, rapid urbanization, and hydrocarbon-intensive economies. Methods developed for temperate regions may not transfer well to this setting.

A specific research need therefore exists: a framework that applies satellite data to estimate national emissions for the GCC, that explicitly quantifies the tradeoff between speed and accuracy, and that produces results transparent enough for policymakers to interpret and use.

1.3 Thesis Objectives

This thesis has four main objectives. First, I assemble a dataset that combines official CO₂ emission records with satellite-derived environmental measurements for all six GCC countries, spanning the period from 2000 to 2023. Second, I develop three distinct XGBoost regression models to compare different data sources: one relies solely on traditional economic indicators (GDP, population, energy consumption) to serve as a benchmark; another uses only satellite observations; and a third merges satellite data with demographic information to create a hybrid approach.

The third objective involves determining which satellite indicators contribute most strongly to emission predictions. I examine whether the patterns identified by the models correspond to the physical and economic realities of the GCC region—specifically, whether they reflect the area's extreme climate, urbanization patterns, and energy-intensive economies. Finally, I quantify the trade-off between timeliness and accuracy. This means measuring how much predictive performance is lost when using satellite data, which become available within days, instead of economic statistics, which typically arrive one to two years after the fact (Global Carbon Project, 2023).

1.4 Thesis Contribution

This work makes several contributions to the literature on satellite-based emission monitoring:

- A novel hybrid framework combining satellite-derived environmental indicators with demographic data for national-level CO₂ emission estimation in the GCC region, achieving R^2 of 0.69 with near-real-time data availability.
- A comprehensive multi-source dataset integrating five satellite products (MODIS LST, MODIS NDVI, DMSP/VIIRS night-time lights, MERRA-2 AOD, and Sentinel-5P NO₂) with official emission statistics for 144 country-year observations.
- An explicit quantification of the accuracy-timeliness trade-off: the satellite-based approach reduces data latency by approximately 99 percent (from 12–24 months to days) at the cost of roughly 27 percentage points in R^2 compared to economic models.
- Physical interpretation of feature importance demonstrates that summer land surface temperature (accounting for 34 percent of model importance) reflects the dominant role of cooling demand in GCC emissions, thereby validating the physical basis of the satellite-emission relationship.
- A reproducible approach based on Google Earth Engine for extracting satellite data and open-source Python tools for modeling, which allows this framework to be adapted for other regions.

1.5 Thesis Organization

The thesis proceeds as follows:

Chapter 2 covers necessary background on greenhouse gas emissions, GCC climate policy, and the satellite data sources used here. It also explains the XGBoost algorithm used for modeling.

Chapter 3 examines past studies on using AI for emission forecasting and satellites for carbon monitoring, highlighting where this work contributes.

Chapter 4 presents the methodology: defining the study area, collecting data, using the Google Earth Engine pipeline to obtain satellite observations, making feature engineering decisions, and designing experiments to compare the three models.

Chapter 5 reports experimental outcomes: model performance measures, feature importance patterns, error analysis, and results broken down by country.

Chapter 6 summarizes key findings, considers practical implications, defines the framework's scope, and closes with main contributions and suggestions for future research.

Chapter 2 Overview

2.1 Introduction

This chapter establishes the conceptual and technical background for the thesis. It opens with a brief discussion of artificial intelligence and machine learning, then focuses on supervised learning and ensemble methods, particularly gradient boosting and XGBoost. Next, it covers essential climate and policy context, including greenhouse gas emissions and the GCC region. The chapter concludes by describing the satellite data sources and their connection to the machine learning approach.

The aim is to show why using satellite data and machine learning to produce near-real-time national CO₂ estimates makes sense both methodologically and practically for the GCC.

2.2 Artificial Intelligence

Artificial Intelligence (AI) refers to a diverse field focused on creating systems capable of tasks that normally require human intelligence: perception, language comprehension, decision-making, and learning from experience (Russell & Norvig, 2020). AI is not one technique but rather includes multiple approaches such as rule-based systems, search and optimization algorithms, and data-driven learning methods.

In environmental research, AI has become a common tool for examining large, varied datasets—satellite images, climate reanalysis products, and socioeconomic indicators, for example (Persello et al., 2022). AI methods can uncover patterns that manual analysis might miss and can produce predictive or diagnostic models linking satellite observations to processes like carbon emissions.

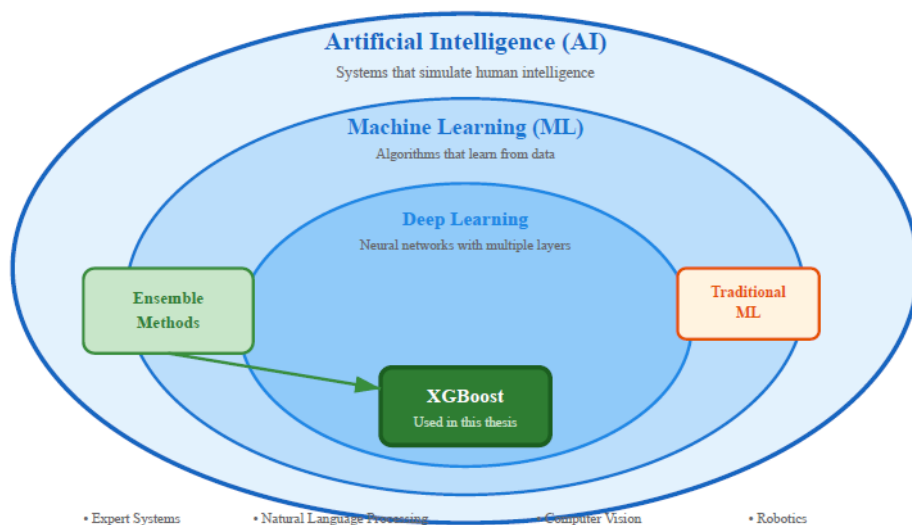


Figure 2.1: Artificial Intelligence and Its Subfields

2.3 Machine Learning

Machine Learning (ML) belongs to AI but operates on different principles. Instead of writing rules explicitly, ML algorithms identify patterns by studying data (Russell & Norvig, 2020). The algorithm examines cases where inputs correspond to known outputs, then builds a function capable of handling inputs it has never seen. This method works well when relationships cannot easily be written as mathematical formulas.

National emissions arise from several interacting factors: population size, economic output, energy use, fuel composition, technological capabilities, and climatic demands. Describing these connections analytically proves challenging. ML addresses this by processing many input variables at once and uncovering how they interact. For this thesis, that means using satellite observations together with socioeconomic indicators to estimate emissions.

This work uses supervised learning exclusively. Training data consist of labeled pairs (x, y) : x contains input features for a particular country and year, while y gives the actual emission total for that case. The algorithm produces a function f where $f(x)$ approximates y , both for training observations and for country-year combinations not used during training.

Other ML techniques exist but are not directly relevant. Unsupervised learning identifies structure in data without labels, useful for exploration rather than prediction. Reinforcement learning handles sequential choices over time, appropriate for control tasks rather than static emission estimation.

Supervised regression matches the research question. Past observations supply known emission values across countries and years. The objective is clear: identify patterns in those records and use them to estimate emissions for periods or places excluded from training. XGBoost, a gradient-boosted tree algorithm, carries out this supervised learning approach and serves as the main modeling tool, explained further in Chapter 4.

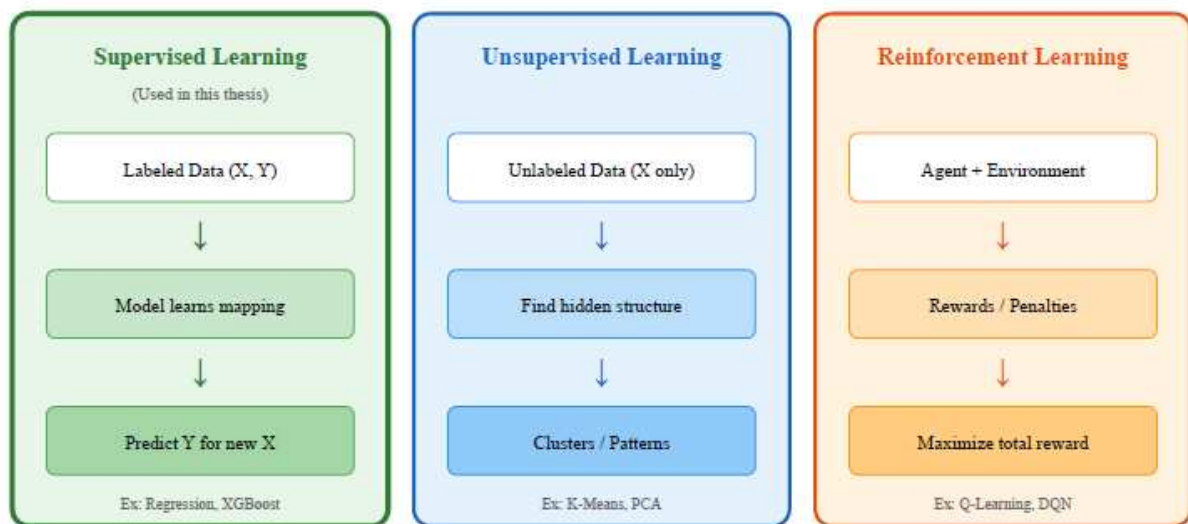


Figure 2.2: Types of Machine Learning

2.4 Supervised Learning in Detail

This thesis relies entirely on supervised learning, so it helps to outline the typical workflow for supervised regression. The process unfolds in several stages.

First comes data collection. This involves assembling a dataset that pairs input features—satellite indicators, population density, basic economic variables—with a target variable, in this case annual national CO₂ emissions, covering historical time periods.

The second stage addresses preprocessing and feature engineering. Missing values need handling, temporal and spatial scales must align, and variables sometimes require transformation (such as taking logarithms). New features derived from raw inputs may also improve how well the model performs.

Third is model training. A regression algorithm gets selected—options include linear models, decision trees, random forests, or gradient-boosted trees—and fitted to a training portion of the data. Training minimizes a loss function, usually squared error for regression tasks.

Fourth, validation and hyperparameter tuning take place. The model's performance gets checked on validation data not used during training. This step allows tuning of hyperparameters and helps prevent overfitting, where the model memorizes training examples rather than learning general patterns.

Finally, testing and evaluation occur. The finished model runs on a held-out test set, and performance metrics get calculated: coefficient of determination (R^2), root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE).

In this thesis, these stages apply to a GCC-wide panel covering multiple countries and years. The data split respects time: models train on earlier years and get tested on later ones, mimicking a real nowcasting scenario where future emissions are unknown.

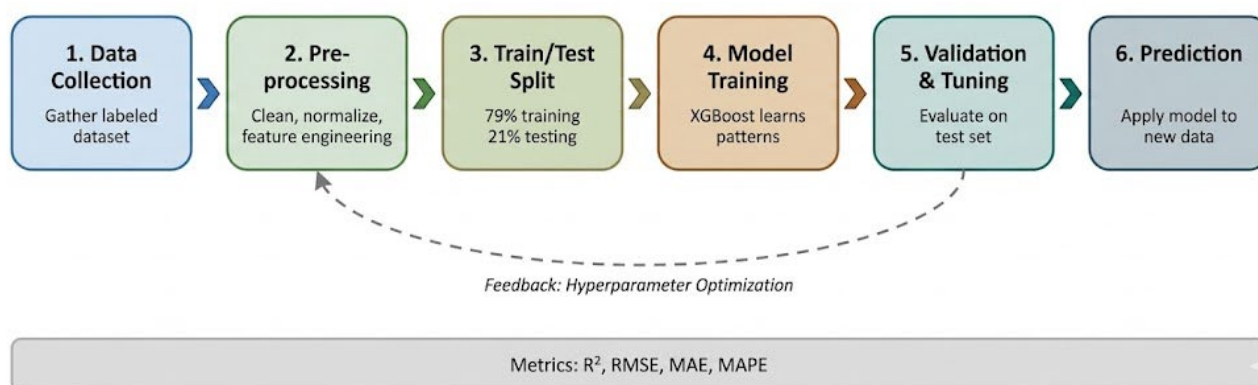


Figure 2.3: Supervised Learning Workflow

2.5 Deep Learning

Deep Learning forms a subset of machine learning built around artificial neural networks containing multiple hidden layers (Goodfellow et al., 2016). These networks comprise interconnected nodes arranged in layers. Each node performs a straightforward mathematical operation: it calculates a weighted sum of its inputs, then applies a nonlinear activation function. Successive layers transform the data progressively into more abstract representations.

Training adjusts the network's weights to reduce prediction error. This typically relies on gradient-based optimization and backpropagation. Deep architectures can represent very complex functions and have set performance records in image recognition, speech processing, and natural language tasks. That said, they usually need large training datasets and considerable computing power. They also contain many parameters, which creates overfitting risk when training examples are scarce. This thesis works with 144 country-year observations—a modest sample. High-capacity deep models struggle under these conditions because they tend to memorize training data instead of extracting stable patterns.

For tabular datasets of small to medium size, gradient-boosted decision trees like XGBoost often perform better than deep neural networks. They deliver strong predictions with fewer parameters and handle limited, heterogeneous data more robustly.

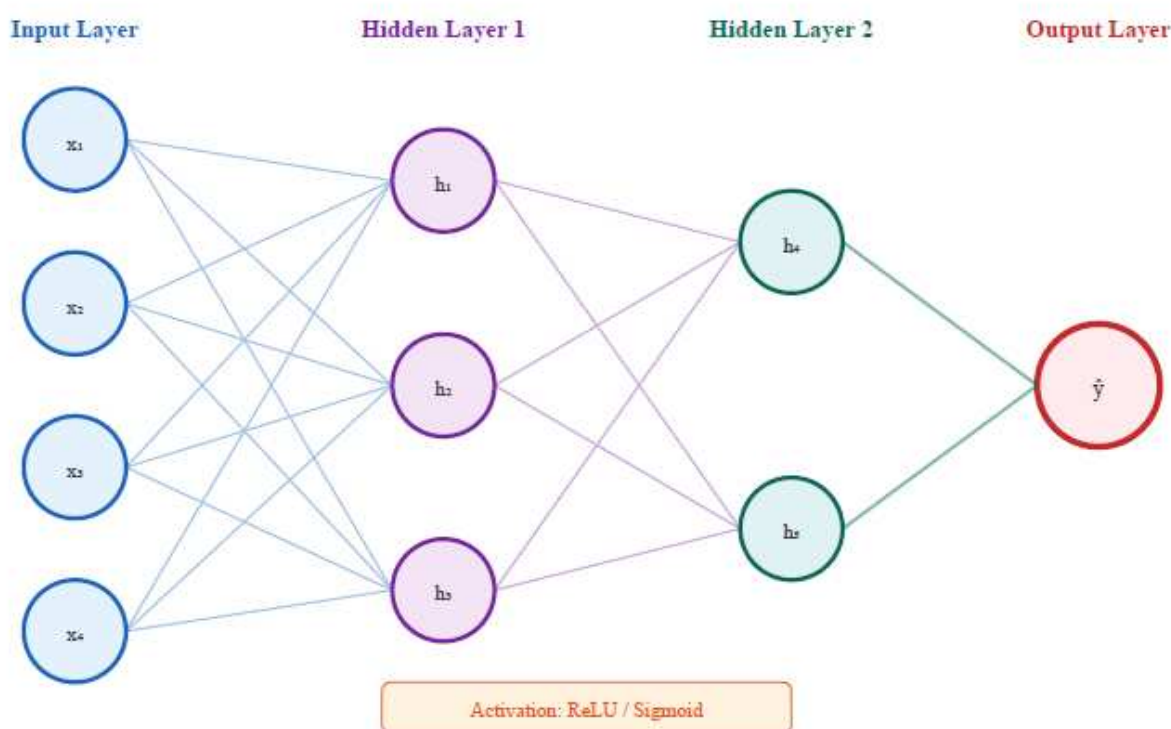


Figure 2.4: Neural Network Architecture

Table 2.1: Comparison of AI, ML, and DL

Aspect	Traditional AI	Machine Learning	Deep Learning
Learning Method	Rule-based, programmed	Learns from data	Learns hierarchical features
Data Requirement	Minimal	Moderate (100s-1000s)	Large (10,000s+)
Interpretability	High	Medium	Low (black box)
Example Algorithms	Expert systems	XGBoost, Random Forest	CNN, Transformers
Best For	Well-defined rules	Tabular data, limited samples	Images, text, large datasets

Source: Author's compilation based on Russell and Norvig (2020) and Goodfellow et al. (2016)

2.6 Ensemble Methods

Ensemble methods combine predictions from multiple base models into a single predictor that usually beats any individual model on both accuracy and robustness (Chen & Guestrin, 2016). The core logic is straightforward: different models pick up different patterns in the data, so merging their outputs tends to reduce both bias and variance.

Three main strategies exist. Bagging—bootstrap aggregating—trains several models independently on bootstrap samples drawn from the training data, then averages their forecasts. Random forests represent a well-known application of bagging to decision trees. Boosting follows a sequential approach: models are constructed one at a time, with each new model targeting the mistakes left by the prior ensemble. The final forecast is a weighted combination of all learners. Gradient boosting and XGBoost exemplify this method. Stacking operates differently: multiple diverse base models train simultaneously, and a separate meta-model figures out how to blend their outputs.

Boosting holds particular relevance here. It performs well on structured, tabular regression problems and handles nonlinear interactions effectively, even when sample sizes are modest.



Figure 2.5: Ensemble Methods Comparison

2.7 The XGBoost Algorithm

XGBoost (Extreme Gradient Boosting) is an efficient, regularized implementation of gradient-boosted decision trees (Chen & Guestrin, 2016). For structured data, it has become a popular choice because it combines strong predictive performance with several practical advantages.

The algorithm works well with different types of features and handles missing values without special treatment. Regularization is built into the training process, which controls model complexity and reduces overfitting risk. Objective functions and evaluation metrics can be customized to fit the problem at hand. Perhaps most useful for this research, XGBoost produces feature importance measures that clarify which variables contribute most to predictions.

In this thesis, XGBoost serves as the core regression method for three model configurations. An economic baseline uses only macroeconomic and demographic indicators. A hybrid model pairs satellite indicators with population density. A satellite-only model draws exclusively on Earth observation proxies.

This choice reflects both the dataset's structure and its limited size. XGBoost provides a reasonable balance between accuracy, robustness, and interpretability.

2.8 Greenhouse Gas Emissions and Climate Policy

Carbon dioxide represents the dominant anthropogenic greenhouse gas. In 2022, fossil fuel combustion and industrial processes released approximately 36.4 billion tonnes of CO₂ globally (Global Carbon Project, 2023). Coal burned for electricity makes up about 40% of this total. Oil consumed in transportation accounts for roughly 25%. Natural gas used for heating and industry contributes another 20%.

Adopted in 2015, the Paris Agreement aims to hold global warming well below 2°C above pre-industrial levels, ideally limiting it to 1.5°C. Achieving this requires steep emission cuts. According to IPCC (2023) estimates, global emissions must decline by approximately 45% from 2010 levels by 2030 (or about 43% from 2019 levels), reaching net zero by 2050.

Countries present Nationally Determined Contributions (NDCs) outlining their emission reduction plans. Assessing progress on these plans needs accurate, current emission data. Most national inventories, however, come out annually with delays of 1–2 years, which makes it hard to judge whether policies are delivering results promptly (Global Carbon Project, 2023).

2.9 The GCC Region: Energy Profile and Climate Commitments

The Gulf Cooperation Council consists of six countries: Saudi Arabia, United Arab Emirates, Qatar, Kuwait, Oman, and Bahrain. Combined, these nations account for roughly 3% of global CO₂ emissions while holding less than 1% of world population. This imbalance stems from their position as leading oil and gas producers and their elevated per capita energy use.

The GCC region experiences a hot, arid climate where summer temperatures regularly surpass 45°C in inland areas. Such conditions create enormous electricity demand for air conditioning, which can reach 60-70% of peak electricity load during summer months. Natural gas generates most of this electricity, creating a tight link between cooling needs and emissions.

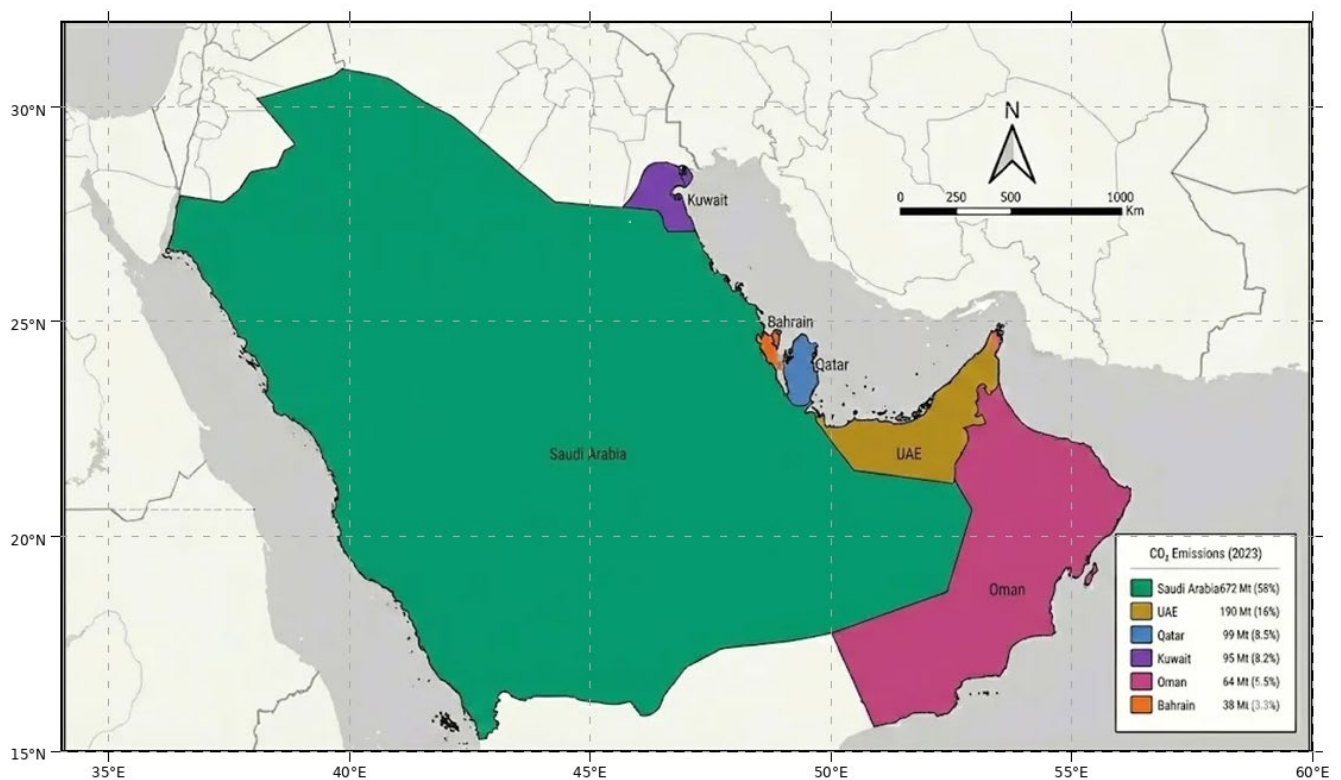


Figure 2.6: Map of the GCC Study Area

Table 2.2: GCC Countries Climate Targets and Emission Profiles

Country	2023 CO ₂ (Mt)	Per Capita (t)	NDC Target	Net Zero Goal
Saudi Arabia	672	18.2	278 Mt reduction by 2030	2060
UAE	190	19.3	31% reduction by 2030	2050
Qatar	99	32.8	25% reduction by 2030	-
Kuwait	95	21.6	7.4% reduction by 2035	-
Oman	64	12.4	7% reduction by 2030	2050
Bahrain	38	23.1	30% reduction by 2035	2060

Source: Global Carbon Project (2023), UNFCCC NDC Registry

All six GCC countries have submitted NDCs, and several have announced net zero targets. Saudi Arabia aims for net zero by 2060, while the UAE and Oman target 2050. These commitments create a need for improved emission monitoring capabilities.

2.10 Earth Observation for Environmental Monitoring

Earth observation satellites have transformed environmental monitoring. Landsat-1, launched in 1972, marked the beginning of an era in which satellite sensors deliver increasingly detailed data on land cover, atmospheric composition, ocean conditions, and climate variables (NASA, 2024).

For monitoring carbon emissions, satellites provide distinct benefits. Coverage is spatially continuous, unlike ground-based networks where gaps exist. Cost barriers fall because many satellite products are freely available, particularly important for developing countries. Data latency is short—satellite observations typically become accessible within days of acquisition, far faster than traditional emission inventories that lag by 1-2 years.

However, satellites measure physical quantities rather than emissions themselves. Surface temperature, night-time lights, and other observable variables must be linked to underlying emission processes. Establishing these relationships reliably forms the central methodological challenge.

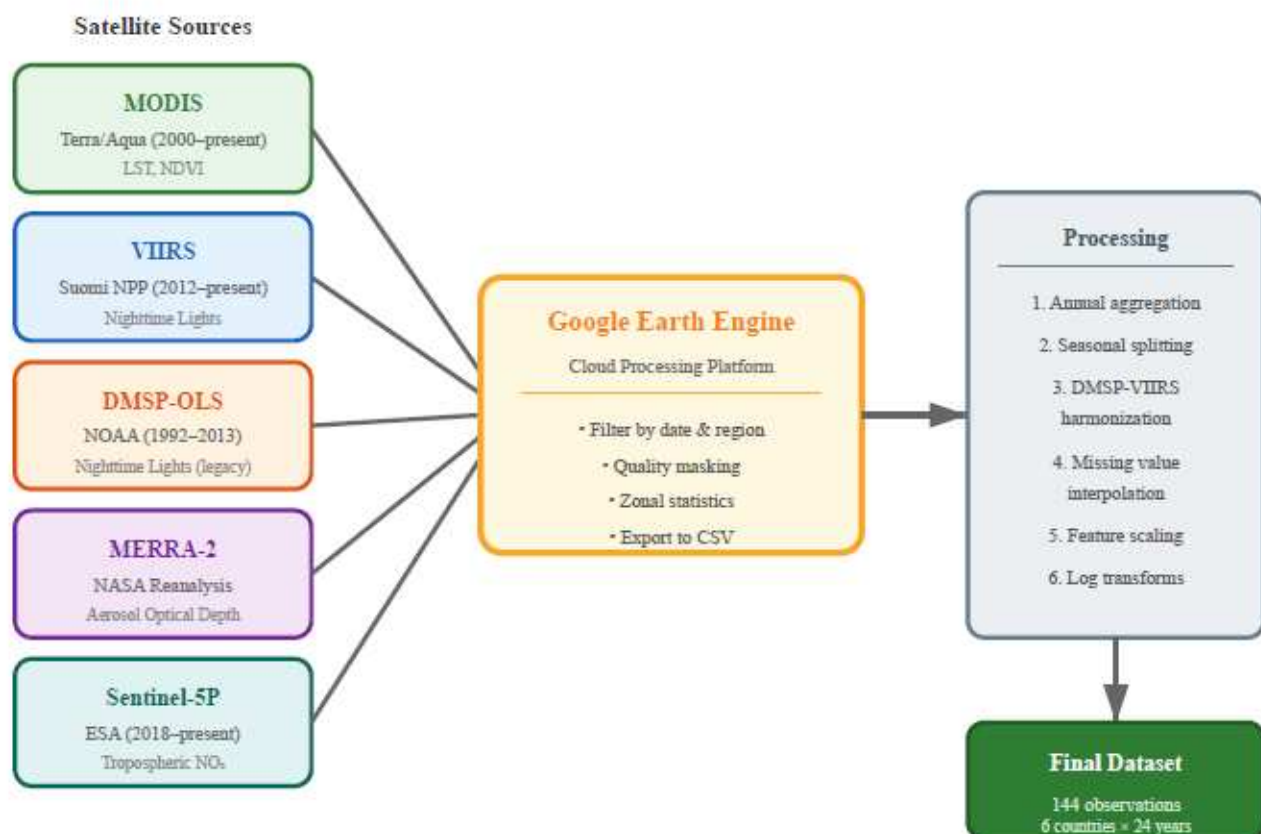


Figure 2.7: Satellite Data Sources and Processing Pipeline

2.11 Satellite Data Sources Used in This Study

This thesis uses five satellite-derived variables, each capturing different aspects of economic activity and energy consumption:

Land Surface Temperature (LST) from MODIS: The Moderate Resolution Imaging Spectroradiometer (MODIS) sensors aboard NASA's Terra and Aqua satellites have been providing LST data since 1999 and 2002 respectively. LST measures the radiative temperature of the Earth's surface using thermal infrared bands. This study utilizes the MOD11A2 product, which provides 8-day composite LST at 1 km resolution. The physical chain linking LST to emissions is straightforward: high temperatures drive air conditioning demand, which increases electricity consumption, which increases emissions. Assiri (2017) validated MODIS thermal data for Saudi Arabia, finding good agreement with ground measurements.

Vegetation Index (NDVI) from MODIS: The Normalized Difference Vegetation Index quantifies vegetation greenness by contrasting near-infrared and red reflectance. Values range from -1 to +1, with values above 0.3 typically indicating healthy vegetation in temperate regions. In the GCC context, natural vegetation is sparse, but irrigated agriculture requires significant energy for water pumping and desalination. The MOD13A2 product provides 16-day NDVI composites at 1 km resolution. Haq (2022) demonstrated 98% classification accuracy for agricultural land cover classification in Al-Qassim, Saudi Arabia, using Planet Scope nanosatellite imagery.

Night-time Lights (NTL) from DMSP-OLS and VIIRS: Night-time light imagery detects artificial lighting from cities, industrial facilities, and infrastructure. Two sensor systems provide this data: the Defense Meteorological Satellite Program (DMSP) Operational Linescan System (OLS) from 1992-2013, and the Visible Infrared Imaging Radiometer Suite (VIIRS) from 2012 onwards. Night-time lights strongly correlate with electricity consumption, GDP, and population density. The main challenge is harmonizing the two sensor records during the 2012-2013 overlap period.

Aerosol Optical Depth (AOD) from MERRA-2: NASA's Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2) provides global reanalysis data including aerosol optical depth. AOD measures the extinction of sunlight by atmospheric particles. In industrial regions, high AOD can indicate combustion emissions. However, the GCC region has naturally high AOD due to desert dust, making interpretation more challenging (Kumar et al., 2018).

Nitrogen Dioxide (NO₂) from Sentinel-5P TROPOMI: The European Space Agency's Sentinel-5P satellite, launched in late 2018, carries the Tropospheric Monitoring Instrument (TROPOMI) which measures NO₂ column density at approximately 5.5 km resolution. NO₂ is a short-lived pollutant concentrated near combustion sources, making it a useful tracer for vehicle and power plant emissions.

High accuracy has been demonstrated in predicting air pollution for Saudi cities using Sentinel-5P data (Alkayal et al., 2025).

Table 2.3: Satellite Sensors and Products Used in This Study

Variable	Sensor/Product	Resolution	Coverage	Physical Meaning
LST	MODIS MOD11A2	1 km, 8-day	2000-present	Thermal emission → cooling demand
NDVI	MODIS MOD13A2	1 km, 16-day	2000-present	Vegetation greenness → agricultural activity
NTL	DMSP-OLS / VIIRS	~750 m	1992-present	Artificial light → urbanization, economic activity
AOD	MERRA-2	~50 km	1980-present	Aerosol loading → combustion + dust
NO ₂	Sentinel-5P TROPOMI	5.5 km, daily	2019-present	Tropospheric NO ₂ → combustion tracer

2.12 Machine Learning for Emission Estimation

Machine learning offers practical tools for identifying links between satellite observations and emissions. Traditional statistical models demand explicit functional forms, whereas machine learning algorithms detect complex, nonlinear patterns automatically from data.

Multiple studies have used machine learning for emission estimation. Nassef et al. (2023) compared Feed-Forward Neural Networks (FFNN), Adaptive Neuro-Fuzzy Inference Systems (ANFIS), and Long Short-Term Memory (LSTM) networks for CO₂ prediction in Saudi Arabia, reaching R² values up to 0.99 with economic predictors. Alamri and Khan (2023) tested N-BEATS and Transformer architectures for emission forecasting, discovering that sophisticated deep learning models did not consistently beat simpler methods when data are scarce.

This thesis selected XGBoost (Extreme Gradient Boosting) for multiple reasons. It processes tabular data efficiently and performs well on structured datasets with limited observations. It generates feature importance scores that clarify which satellite indicators matter most for predictions. It runs efficiently without requiring GPU hardware. Compared to deep learning approaches, it has fewer hyperparameters, which reduces overfitting risk when sample sizes are modest.

2.13 XGBoost: Mathematical Formulation

XGBoost, introduced by Chen and Guestrin (2016), is an optimized implementation of gradient boosted decision trees. The key idea is to build an ensemble of weak learners (decision trees) sequentially, where each new tree corrects the errors of the previous ensemble.

The algorithm minimizes a regularized objective function:

$$L(\theta) = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k) \quad (2.1)$$

where l is a differentiable loss function and Ω is a regularization term

The regularization term Ω penalizes model complexity by including the number of leaves and the magnitude of leaf weights. This helps prevent overfitting, which is particularly important when working with limited data.

For regression tasks like emission prediction, XGBoost typically uses squared error loss. Each tree is fit to the negative gradient of the loss (the residuals from the current ensemble), and predictions are made by summing the outputs of all trees.

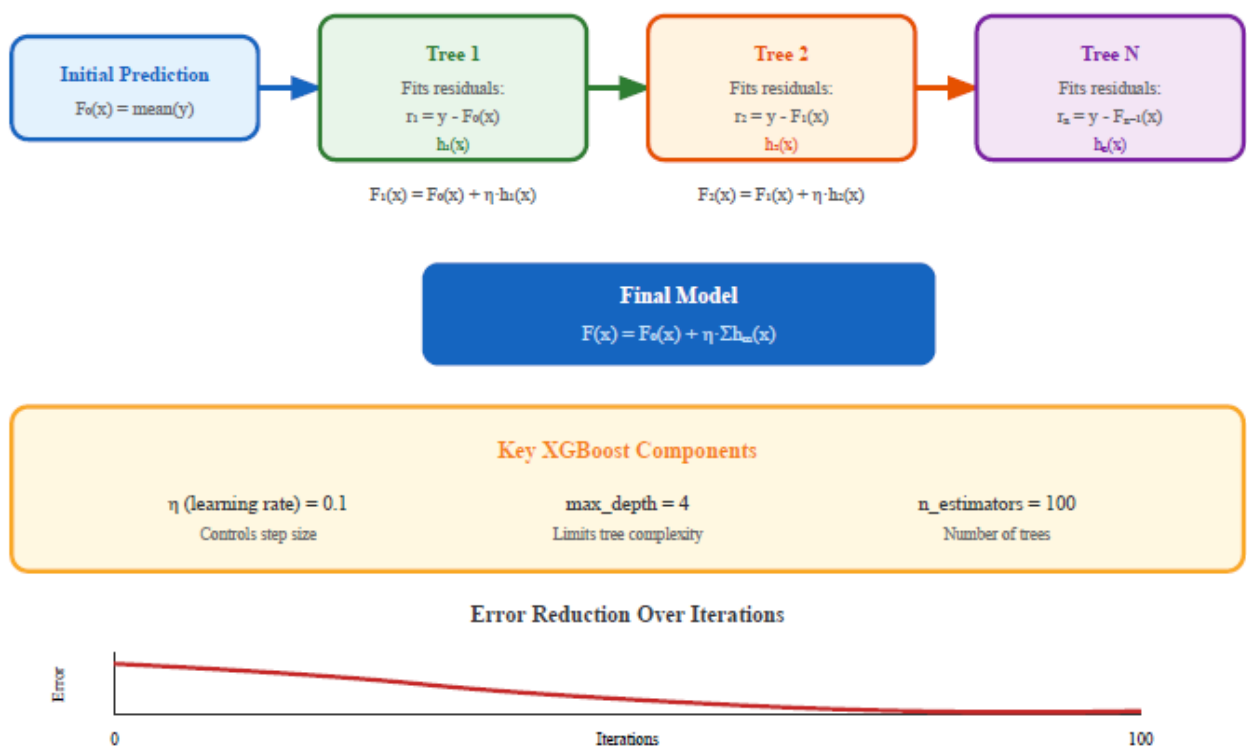


Figure 2.8: XGBoost Sequential Tree Building Process

Key hyperparameters include `n_estimators` (number of trees), `max_depth` (maximum tree depth), `learning_rate` (shrinkage factor applied to each tree), and `min_child_weight` (minimum sum of instance weights in a leaf). These parameters control the trade-off between model expressiveness and overfitting risk.

2.14 Explainable AI for Tree-Based Models

Explainable artificial intelligence (XAI) has become an important addition to purely predictive machine learning models in recent years. The core principle holds that models supporting policy or high-stakes decisions must be accurate and also transparent enough for domain experts to grasp why particular predictions emerge. For environmental applications, interpretability carries special value: it confirms that models depend on physically sound relationships rather than spurious correlations, and it enables policymakers to trace how shifts in factors like energy use, urbanization, or climate conditions influence estimated emissions.

XGBoost and similar tree-based ensemble methods provide multiple explanation strategies. Global feature importance measures summarize each input variable's contribution to overall error reduction across all ensemble trees. These scores let you rank variables and check whether the strongest predictors make sense given what's known about the domain. Local methods, including Shapley-value-based explanations, can attribute individual predictions to specific feature contributions. However, they demand extra computation and prove more vulnerable to sampling noise when datasets are small.

This thesis works with 144 country-year observations containing mixed features. Given this size and structure, the analysis emphasizes global explanations from XGBoost feature importance. This choice offers a practical compromise: it shows whether satellite indicators like land surface temperature or night-time lights behave as anticipated when aggregated, without over-interpreting noisy instance-level patterns that small datasets cannot reliably support. Local XAI methods remain a direction for future research rather than forming part of the current analysis.

Chapter 3 Related Work

3.1 Introduction

This chapter surveys prior research where artificial intelligence meets carbon emissions monitoring. Two areas matter here: work that uses machine learning to forecast emissions, and work that uses satellites to track atmospheric carbon. Knowing what these studies achieved—and what they missed—clarifies this thesis's contribution.

Publications from 2021 through 2025 were reviewed, guided by three questions: Which AI techniques have researchers applied to CO₂ prediction? How are satellites being used to monitor emissions? What studies exist for Gulf settings specifically? Thirteen studies proved most relevant. Seven applied machine learning to forecasting, six focused on satellite monitoring. A pattern emerges across these papers: researchers are pairing remote sensing with AI more often, though gaps persist in how fast these methods produce usable estimates.

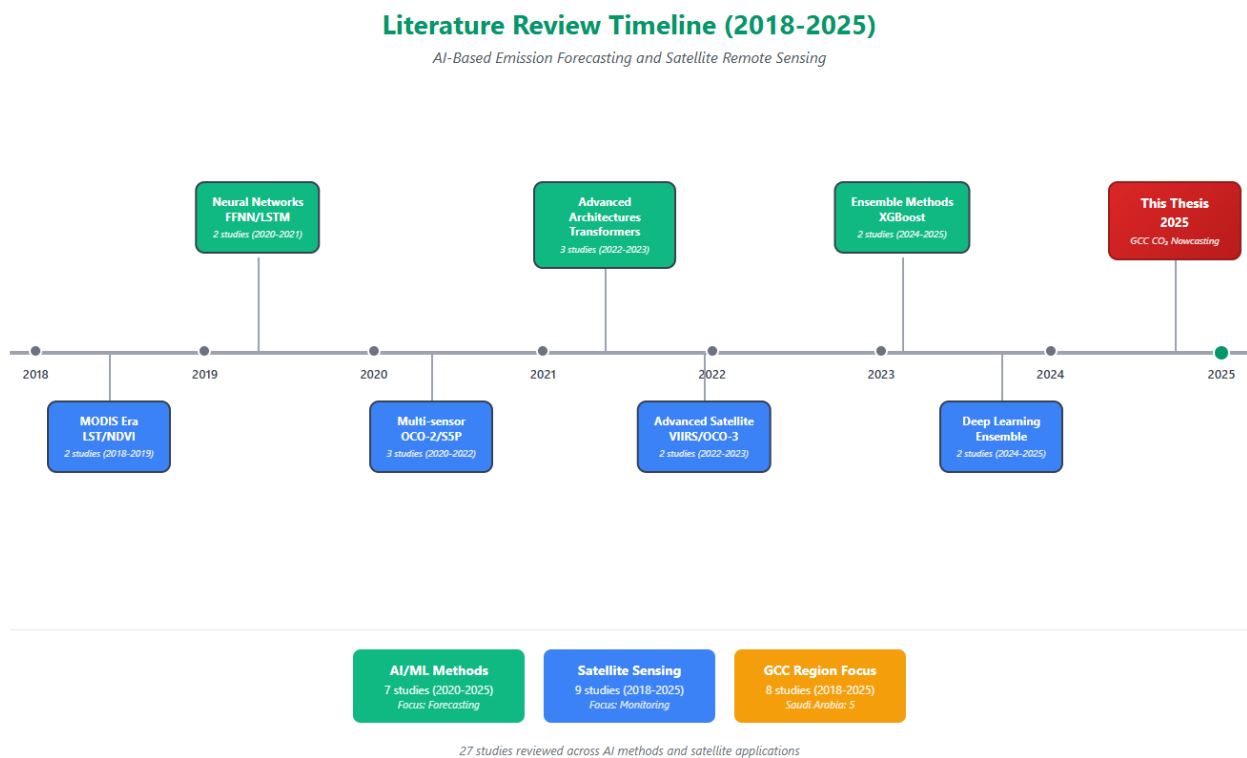


Figure 3.1: Literature Review Timeline (2016-2025)

3.2 AI-Based Emission Forecasting

Machine learning for carbon prediction has grown substantially in recent years. Early studies used basic regression, while newer work has explored more elaborate techniques.

Nassef et al. (2023) created a forecasting system for Saudi CO₂ emissions using three algorithms: FFNN, ANFIS, and LSTM. They worked with economic inputs like GDP, population, energy consumption, and vehicle numbers, achieving R² values from 0.97 to 0.99. FFNN gave the best performance, suggesting simpler models can work better than complex ones when data is limited. Still, their approach uses economic statistics that arrive with the same one-to-two-year lag this thesis tries to avoid.

Rahman et al. (2023) employed Decision Tree Regression for CO₂ forecasting from fuel consumption. Their model hit R² = 0.99 on historical records, though the authors acknowledged sensitivity to fuel data gaps and quality issues. Meng and Noman (2022) studied COVID-19's impact on global CO₂ through machine learning, observing that pandemic shocks created patterns conventional forecasting struggles to capture.

Alamri and Khan (2023) evaluated N-BEATS and Transformer models for emission forecasting. Despite advantages for sequential data in theory, these deep learning approaches couldn't consistently beat simpler baselines—probably due to limited training samples. This supports using XGBoost in this thesis rather than deep learning.

Alcibahy et al. (2025) recently demonstrated XGBoost's success in downscaling satellite CO₂ observations over the Arabian Peninsula. Their ensemble combined MODIS LST, OCO-2, Sentinel-5P, and ERA-5 data, reaching strong performance (R² = 0.98, RMSE = 0.58 ppm) at 1 km resolution. This validates combining gradient boosting with multiple satellite sources in GCC arid conditions.

Begum and Al Mobin (2025) analyzed ensemble ML methods for emission forecasting across 11 major emitters, including Saudi Arabia. They used XGBoost, SVR, and Gradient Boosting with Sequential Squeeze Feature Selection, achieving 96.21% average accuracy. Their projections show Saudi emissions might climb 16.62% by 2030 rather than decline, emphasizing the urgency of improved monitoring.

Han et al. (2023) tested hybrid LSTM-CNN structures for provincial Chinese emissions, obtaining R² = 0.97 with spatial correlation weighting. Deep learning works well with large datasets. However, gradient boosting like XGBoost fits small temporal datasets (144 observations in this study) better—it has lower overfitting risk and easier interpretation.

Fonseca and Francis (2023) documented baseline CO₂ patterns across the Middle East from nine years of OCO-2/OCO-3 observations (2014-2023). Seasonal XCO₂ variations of 3-8 ppmv appeared over the Arabian Peninsula, with UAE trends at 2.50 ppmv/year. These regional baselines provide validation benchmarks for GCC forecasts.

Xu et al. (2023) showed NPP-VIIRS nighttime light data serves as an effective proxy for human CO₂ emissions at fine spatial scales. This approach, built into the proposed framework through VIIRS DNB data, complements direct atmospheric CO₂ measurements for locating emission sources in urban GCC regions.

Table 3.1: Summary of AI-Based Emission Forecasting Studies

Authors	Method	Region	Best R ²	Data Type
Nassef et al. (2023)	FFNN, ANFIS, LSTM	Saudi Arabia	0.99	Economic
Rahman et al. (2023)	Bagged DT	Saudi Arabia	0.90	Economic
Alamri and Khan (2023)	N-BEATS, Transformer	Multiple	0.95	Economic
Meng and Noman (2022)	SARIMAX	Global	N/A (MAPE: 0.09)	Economic
Alcibahy et al. (2025)	XGBoost, RF, SVM	Arabian Peninsula	0.98	MODIS, OCO-2, S5P
Begum and Al Mobin (2025)	XGBoost, SVR, GB	Saudi Arabia	96.21%	Economic data
Han et al. (2023)	LSTM-CNN	China	0.97	Historical

3.3 Satellite-Based Carbon Monitoring

Satellite-based carbon research follows two tracks. Some studies measure atmospheric CO₂ directly. Others estimate emissions indirectly through proxy variables.

Direct measurement uses specialized satellites like NASA's OCO-2 (Orbiting Carbon Observatory-2). Mustafa et al. (2021) estimated regional anthropogenic CO₂ over East and West Asia with neural networks trained on OCO-2 data. Alcibahy et al. (2025) applied this to the Arabian Peninsula, combining OCO-2 CO₂ with Sentinel-5P methane data. Their XGBoost model achieved strong performance, demonstrating multi-sensor fusion benefits.

Indirect approaches use satellite observations as proxies for emission activities. Lin et al. (2023) developed machine learning to estimate power plant CO₂ from satellite plume observations. Rowley and Karakuş (2023) created AQNet, predicting air quality from satellite imagery. Persello et al. (2022) reviewed deep learning for sustainable development goals, noting satellite data potential for environmental monitoring.

Land applications include Mallick et al. (2022), who used remote sensing for carbon stock estimation in Saudi landscapes. Madugundu et al. (2018) combined remote sensing with soil sampling for agricultural carbon assessment. These studies show satellite-based carbon evaluation is feasible in the GCC, though they focus on carbon stocks rather than emissions.

Table 3.2: Summary of Satellite-Based Carbon Monitoring Studies

Authors	Satellite Data	Region	Approach	Application
Mustafa et al. (2021)	OCO-2	East/West Asia	Neural Network	CO ₂ estimation
Alcibahy et al. (2025)	OCO-2, Sentinel-5P	Arabian Peninsula	XGBoost	CO ₂ , CH ₄ estimation
Lin et al. (2023)	Multiple	Global	ML ensemble	Power plant emissions
Rowley and Karakuş (2023)	Multiple	Urban areas	Deep learning	Air quality
Fonseca and Francis (2023)	OCO-2, OCO-3	Middle East	Trend analysis	Regional variability
Xu et al. (2023)	NPP-VIIRS NTL	China	Nighttime light	Urban emissions

3.4 Regional Studies in the GCC

Environmental monitoring in the GCC has been studied in several works, though few focus specifically on CO₂ emission estimation.

Assiri (2017) validated MODIS thermal data for Saudi Arabia's land surface temperature analysis. Ground measurements matched well across different land cover types. This supports MODIS LST as a reliable thermal proxy in the region.

Haq (2022) reached 98% accuracy classifying agricultural vegetation in Al-Qassim using Planet Scope nanosatellite imagery with machine learning. Addas et al. (2020) studied urban heat islands in Jeddah through satellite data, observing major temperature differences between urban and surrounding zones.

Hamam and Mohamed (2020) mapped soil salinity in the East Nile Delta using multiple salinity assessment approaches. Alkayal et al. (2025) demonstrated high accuracy predicting air pollution for three Saudi cities with Sentinel-5P data and deep learning.

Carbon-related work includes Emadi et al. (2020), who used stacked machine learning models for soil property prediction in arid areas, proving ensemble methods work. Tan et al. (2024) applied K-means clustering to group Chinese regions by emission traits, showing that regional clustering combined with satellite XCO₂ data boosted carbon estimation accuracy.

3.5 Research Gaps and Contributions

The literature review shows several gaps this thesis addresses:

- Most AI forecasting for emissions uses economic indicators that arrive with 1–2-year delays. Satellite data as an alternative input has received little attention.
- Satellite carbon monitoring typically targets global scales or single facilities. National-level estimation for regions like the GCC remains underdeveloped.
- The accuracy-timeliness trade-off has not been measured explicitly. Studies report high accuracy using economic data but skip comparisons with satellite alternatives.
- Regional validation for GCC climate conditions is scarce. Most satellite-emission relationships come from temperate regions and may not apply directly here.

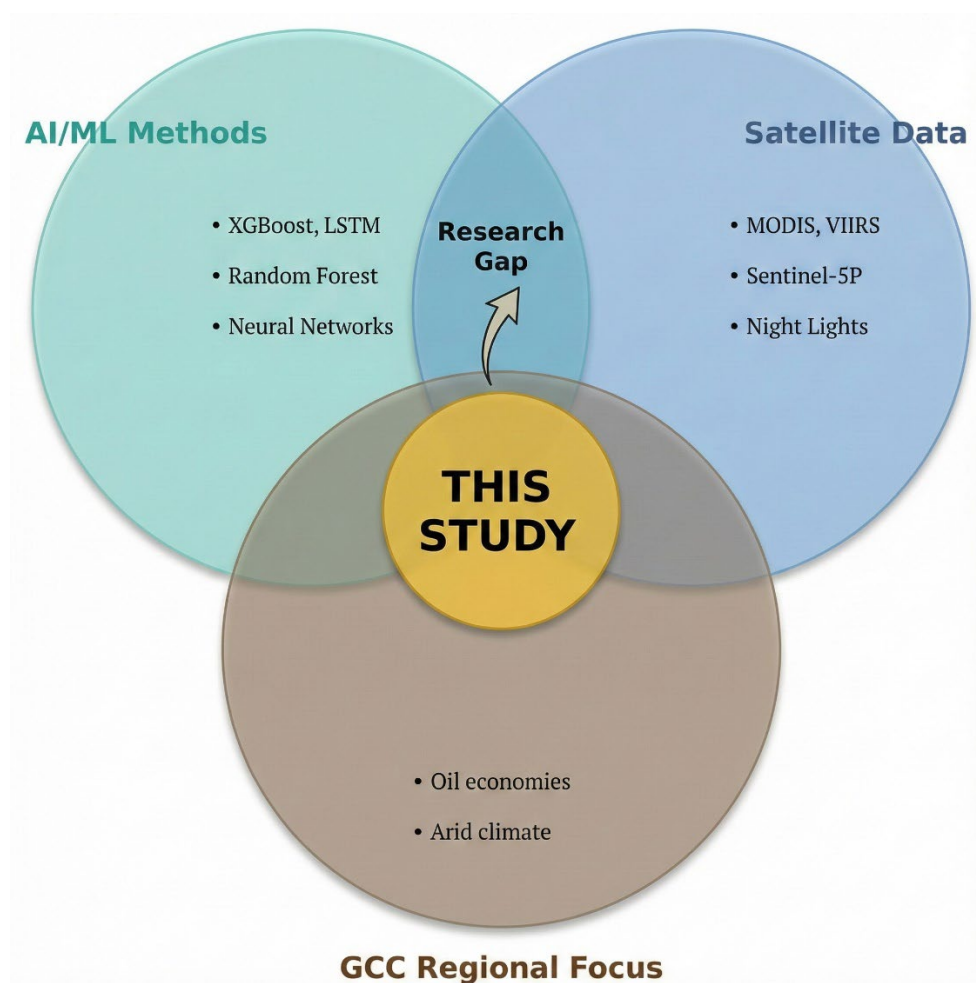


Figure 3.2: Research Gap Analysis Diagram

This thesis helps fill these gaps through three contributions. It develops a satellite framework validated specifically for the GCC region. It compares economic and satellite approaches directly to measure the accuracy-timeliness trade-off. It provides interpretable feature importance analysis that confirms the physical basis of satellite-emission relationships in hot, arid climates.

Chapter 4 Proposed Approach

4.1 Introduction

This chapter describes the methodological framework built to estimate national CO₂ emissions for Gulf Cooperation Council (GCC) countries using satellite-derived indicators. The approach combines multi-sensor satellite observations with machine learning to generate near-real-time emission estimates that supplement conventional inventory-based statistics.

The framework addresses one question: how well can satellite data, available within days to weeks, approximate official national emission totals that typically arrive only after 12–24 months? The chapter covers dataset construction, model design, and the evaluation strategy used to measure the accuracy-timeliness trade-off.

4.2 Research Design Overview

The study uses a quantitative, comparative design with three components.

First, a unified panel dataset was built combining official national CO₂ emission statistics with satellite-derived environmental indicators for all six GCC countries over 2000–2023. Emission time series come from the Global Carbon Project, compiled by Our World in Data (2023), aligned with annual aggregates from multiple satellite products.

Second, three regression models based on Extreme Gradient Boosting (XGBoost) (Chen & Guestrin, 2016) were trained and compared: an economic baseline using conventional economic and energy indicators (Model A); a hybrid combining satellite indicators with population density (Model B); and a satellite-only model using remotely sensed variables alone (Model C).

Third, model performance, feature importance, and error patterns were analyzed on a held-out test period (2019–2023) to characterize satellite-based estimation strengths and limits in the GCC context.

Before settling on the final design, preliminary experiments explored how far back the analysis could extend and whether single-country or multi-country setups worked better. Using OWID/Global Carbon Project and World Bank statistics, exploratory models tested longer historical windows reaching into the

1980s and 1990s, focusing only on Saudi Arabia. These longer windows increased sample size but exposed problems: reporting conventions shift in early years, documentation for some economic indicators is incomplete, and resulting models became more sensitive to train-test splits. Models calibrated on very early years showed high in-sample fit but unstable performance on recent years.

To avoid conclusions based on unstable historical segments, the final study restricts analysis to 2000–2023. From 2000 onwards, emission and economic series are better documented and most satellite products used here are consistently available. MODIS LST and NDVI start around 2000, VIIRS night-time lights become usable after inter-calibration with DMSP-OLS, and MERRA-2 aerosol products are fully available over the study region. Sentinel-5P NO₂ is only available from late 2018 onwards, so it's treated as an auxiliary variable with partial coverage rather than a core driver. The 2000–2023 window still yields 144 country-year observations across six GCC countries, balancing data quality, temporal coverage, and model complexity.

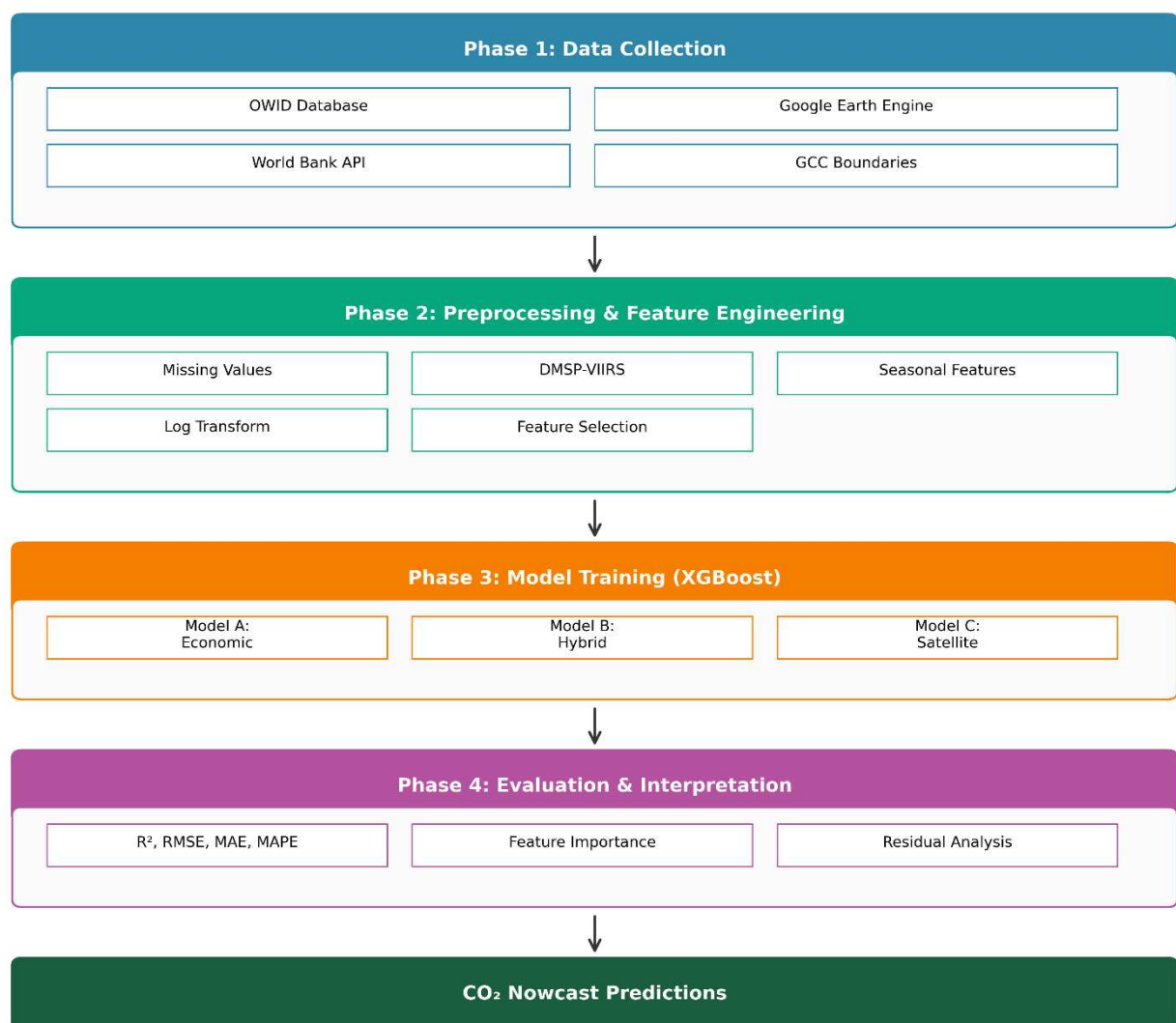


Figure 4.1: Research Framework Architecture

The choice of XGBoost (Chen & Guestrin, 2016) over deep learning architectures is deliberate. With only 144 observations ($6 \text{ countries} \times 24 \text{ years}$), the dataset is too small for data-hungry deep learning models without incurring substantial overfitting risk. XGBoost provides built-in regularization, handles non-linear relationships and interactions effectively, and produces feature-important measures that support interpretation. These properties make it suitable for tabular datasets of limited size.

From a modelling perspective, the 144 observations are treated as a pooled country–year panel. A single XGBoost model is estimated across all six GCC countries, without country-specific fixed or random effects. The objective is purely predictive – to generate accurate nowcasts of national CO₂ emissions – rather than to identify causal, country-level parameter estimates.

4.3 Study Area: The GCC Region

The study covers all six Gulf Cooperation Council countries: Saudi Arabia, United Arab Emirates (UAE), Qatar, Kuwait, Oman, and Bahrain. These countries occupy roughly 2.57 million km² of the Arabian Peninsula and had a combined population of about 58.8 million in 2023.

Table 4.1 summarizes key study area characteristics: land area, population, estimated CO₂ emissions, and typical summer temperature ranges.

Table 4.1: Study Area Overview - Six GCC Countries

Country	Area (km ²)	Population (M)	2023 CO ₂ (Mt)	Climate	Summer Temp (°C)
Saudi Arabia	2,149,690	36.4	672	Arid/Semi-arid	42-50
UAE	83,600	9.4	190	Arid	40-48
Qatar	11,586	2.7	99	Arid	40-46
Kuwait	17,818	4.3	95	Arid	42-50
Oman	309,500	4.5	64	Arid/Tropical	38-45
Bahrain	780	1.5	38	Arid	38-42
GCC Total	2,572,974	58.8	1,158	-	-

Source: World Bank, Global Carbon Project (2023), national statistics offices

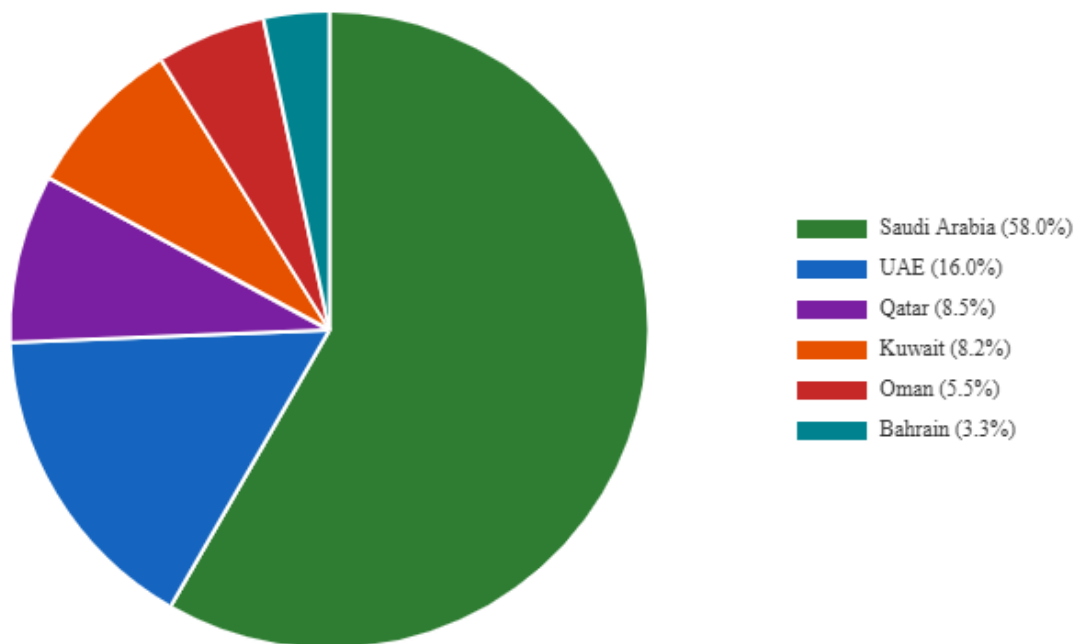


Figure 4.2: GCC CO₂ Emissions Distribution by Country (2023)

The GCC region has a predominantly hot, arid climate. Summer temperatures often exceed 45°C in interior areas, creating a pronounced seasonal cycle in electricity demand driven mostly by air conditioning. In many GCC systems, summer peak electricity demand reaches about 2–3 times winter demand. This seasonal pattern appears in emissions from power generation. The proposed framework exploits this link between thermal conditions and emissions through seasonal land surface temperature (LST) features.

4.4 Data Sources and Collection

The empirical analysis uses two main data streams: national CO₂ emission records and environmental indicators from economic statistics and satellite observations.

4.4.1 Emission Data

Annual national CO₂ emission data came from Our World in Data (Our World in Data, 2023), which compiles statistics from the Global Carbon Project (2023). In this thesis, the target variable is national CO₂ emissions for each GCC country over 1965–2023, reported in million tons (Mt) of CO₂. OWID/GCP also provide longer historical series for some countries, but I rely on the harmonized panel used in this study, and the main analysis is reported for the 2000–2023 period to match satellite data availability.

The harmonized emission series was stored in a pre-processed GCC emissions dataset. Though underlying data are available from 1965 onwards, the empirical analysis here is restricted to 2000–2023 to ensure consistency with satellite product temporal coverage.

4.4.2 Economic and Demographic Data

For the economic baseline model (Model A), these indicators were extracted from the same OWID database, which aggregates World Bank statistical sources:

- Gross Domestic Product (GDP) in constant 2015 USD
- Population (annual estimates)
- Primary energy consumption (terawatt-hours)
- Energy per capita (primary energy divided by population)

These indicators are stored in the same `gcc_emissions_preprocessed_1965_2023.csv` file, with consistent country and year keys, facilitating direct merging with the emission time series.

4.4.3 Satellite Data

Satellite observations were extracted using Google Earth Engine (GEE), which provides access to multiple global image collections, including MODIS, VIIRS, MERRA-2, and Sentinel-5P (Assiri, 2017; Elvidge et al., 2009; European Space Agency, 2024).

After processing, aggregated satellite indicators were stored in a separate seasonal satellite dataset, later merged with emissions and economic data to form the final analysis panel.

Table 4.2 summarizes the main data sources and their temporal coverage.

Table 4.2: Data Sources and Temporal Coverage

Data Type	Source	Coverage	File Name
CO ₂ Emissions	Our World in Data / Global Carbon Project (2023)	1965–2023	Pre-processed GCC emissions dataset
Economic Indicators	Our World in Data / World Bank	1965–2023	Pre-processed GCC emissions dataset
Satellite Features	Google Earth Engine (multiple products)	2000–2023	Seasonal satellite indicators dataset

After merging emission and satellite files on the (country, year) key, the final analysis dataset contains 144 observations (6 countries × 24 years).

4.5 Satellite Data Extraction Pipeline

The satellite data extraction pipeline was built in Google Earth Engine (Gorelick et al., 2017). GEE organizes imagery into Image Collections—time series of images with consistent metadata. For each variable, a collection was selected along with band selection, scale factors, and quality masks.

4.5.1 Google Earth Engine Image Collections

Table 4.3 lists main GEE collections used here, with bands and scale factors applied. MODIS land surface temperature products, validated for Saudi Arabia (Assiri, 2017), provide core thermal indicators. VIIRS night-time lights, MERRA-2 aerosol optical depth, and Sentinel-5P NO₂ capture anthropogenic activity and atmospheric composition (Elvidge et al., 2009; European Space Agency, 2024).

Table 4.3: Google Earth Engine Image Collections Used

Variable	GEE Collection ID	Band Used	Scale Factor	Unit
LST Day	MODIS/061/MOD11A2	LST_Day_1km	0.02	Kelvin
LST Night	MODIS/061/MOD11A2	LST_Night_1km	0.02	Kelvin
NDVI	MODIS/061/MOD13A2	NDVI	0.0001	Index (-1 to 1)
NTL	NOAA/VIIRS/DNB/MONTHLY_V1/CMSELCFG	avg_rad	1.0	nW/cm ² /sr
AOD	NASA/GSFC/MERRA/aer/2	TOTEXTTAU	1.0	Dimensionless
NO ₂ Signal	COPERNICUS/S5P/OFFL/L3_NO2	tropospheric NO2 column number density	1.0	mol/m ²

Note: DMSP-OLS data (1992–2013) were handled separately and inter-calibrated with VIIRS night-time lights (Elvidge et al., 2009).

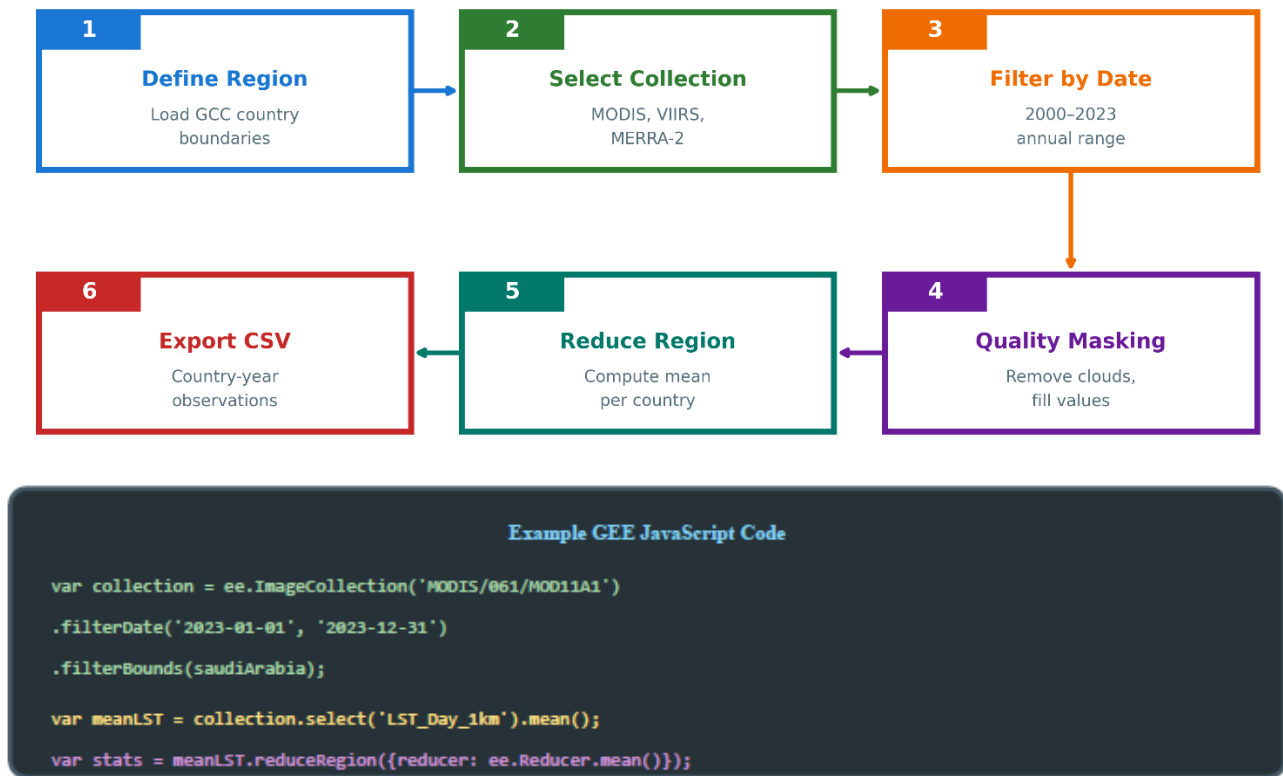


Figure 4.3: Google Earth Engine Data Extraction Workflow

4.5.2 Extraction Workflow

For each country and year, the following steps were applied in GEE:

- **Step 1 – Define Country Boundaries:** Country boundaries were taken from the FAO GAUL dataset available in GEE. Each GCC country is represented as a polygon feature.
- **Step 2 – Filter Image Collection:** For each variable, the relevant Image Collection was filtered by spatial extent (country polygon) and temporal window. Annual variables use the full calendar year (1 January to 31 December), whereas seasonal variables use specific month ranges.
- **Step 3 – Apply Quality Filters:** Quality assurance (QA) flags were used to mask low-quality pixels. For MODIS LST, pixels flagged as cloud-contaminated or low-accuracy in the LST_QC band were excluded. For Sentinel-5P NO₂, only pixels with qa_value > 0.5 were retained (European Space Agency, 2024).
- **Step 4 – Temporal Aggregation:** After QA masking, the collection was reduced temporally using the mean reducer to produce a single composite image per country-year (or country-season).
- **Step 5 – Spatial Aggregation:** The composite image was then reduced over the country polygon using reduce Region with a mean reducer, yielding one value per variable, per country, per year.

- **Step 6** – Export Results: The aggregated values were exported to a CSV file with columns for country, year, and each satellite feature.

Table 4.4 summarizes the most important extraction parameters.

Table 4.4: Satellite Feature Extraction Parameters

Variable	Temporal Filter	Quality Mask	Reducer	Output
LST_Summer	June-August	LST_QC < 65	Mean	°C (converted)
LST_Winter	Dec-Feb	LST_QC < 65	Mean	°C (converted)
NDVI_Spring	March-May	Pixel reliability ≤ 1	Mean	Index
NTL_Annual	Full year	$0 < \text{radiance} < 1000$	Mean	nW/cm ² /sr
AOD_Annual	Full year	Valid flag	Mean	Dimensionless
NO2_Signal	Full year	qa_value > 0.5	Mean	mol/m ²

4.5.3 Handling the DMSP–VIIRS Transition

Night-time lights are observed by two different systems. The DMSP-OLS series provides yearly composites for 1992–2013, while VIIRS offers monthly radiance data from 2012 onwards (Elvidge et al., 2009). To obtain a consistent time series for 2000–2023, an inter-calibration procedure was applied:

- Estimate a regression relationship between DMSP digital numbers and VIIRS radiance during the overlap period 2012–2013.
- Use the calibration coefficients to convert all pre-2012 DMSP values into VIIRS-equivalent radiance.
- Construct a continuous series using calibrated DMSP for 2000–2011, the average of both sensors for 2012–2013, and VIIRS alone from 2014 onwards.

4.5.4 Missing Data Handling

Satellite records are subject to data gaps arising from sensor issues, cloud cover, and incomplete temporal coverage. The Sentinel-5P NO₂ product is particularly affected because the mission began only in late 2018 (European Space Agency, 2024).

Table 4.5: Missing Data Summary

Variable	Missing Count	Missing %	Reason
NO2_Signal	114	79.2%	Sentinel-5P available only after 2018
primary_energy_consumption	1	0.7%	Reporting gap
All other variables	0	0%	—

NO₂ from Sentinel-5P was excluded from the final model due to its limited temporal coverage (only 2019 onwards).

4.6 Feature Engineering and Preprocessing

Feature engineering focused on aligning the satellite indicators with emission-relevant processes in the GCC and on stabilizing the statistical properties of the variables before model training.

4.6.1 Seasonal Disaggregation of Temperature

Initial tests using annual mean LST exhibited relatively weak correlation with CO₂ emissions. In the GCC, emissions from electricity generation are driven primarily by cooling demand in the hot season rather than by heating in winter. To capture this asymmetry, LST was disaggregated into seasonal components: LST_Summer_C (mean land surface temperature over June–August) and LST_Winter_C (mean land surface temperature over December–February). In addition, NDVI was aggregated over March–May to capture the main vegetation growth period and associated agricultural activity. Table 4.6 summarizes the aggregation windows and their rationale.

Table 4.6: Seasonal Aggregation Windows

Feature	Months	Rationale
LST_Summer_C	June–August	Peak cooling season; temperatures often 45–50°C; drives AC demand
LST_Winter_C	December–February	Baseline thermal conditions; minimal heating in GCC
NDVI_Spring	March–May	Vegetation growth period; captures agricultural activity

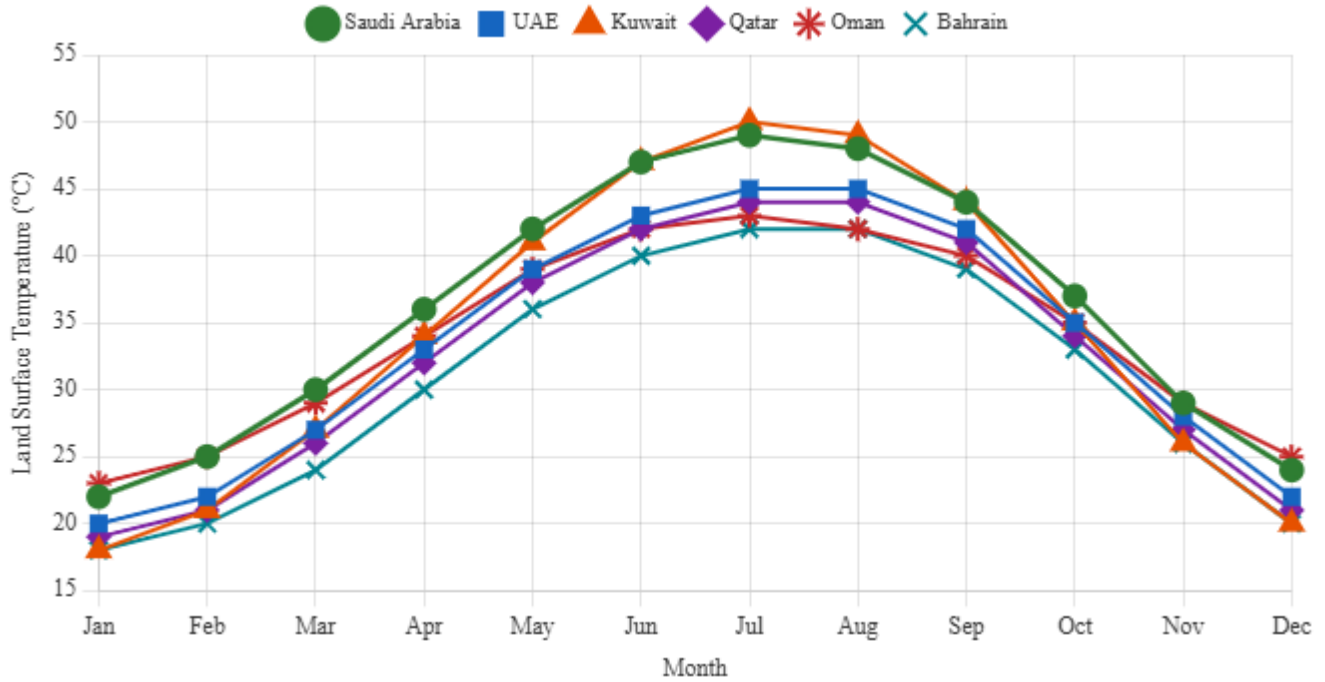


Figure 4.4: Seasonal Land Surface Temperature Pattern

4.6.2 Population Density Transformation

Population density varies substantially across the GCC, from roughly 12 people/km² in Oman to more than 1,700 people/km² in Bahrain. This high skew can affect tree-based models and error metrics. To stabilize the distribution, population density was transformed using a base-10 logarithm:

$$\log_pop_density = \log_{10}(\text{population} / \text{land_area}) \quad (4.1)$$

This transformation converts multiplicative differences into additive ones and reduces the impact of extremely dense observations on the model.

4.6.3 Final Feature Set

After feature engineering, the models use the input variables listed in Table 4.7. Model A represents a purely economic baseline. Model B combines satellite variables with population density. Model C uses only satellite features to assess the standalone predictive power of remote sensing.

Table 4.7: Final Feature Set for Each Model

Model	Feature Category	Variables
Model A (Economic)	Economic Demographic	+GDP, population, energy_consumption, energy_per_capita
Model B (Hybrid)	Satellite Demographic	+LST_Summer_C, LST_Winter_C, NDVI_Spring, Night_Lights, AOD_Annual, log_pop_density
Model C (Satellite-only)	Satellite Only	LST_Summer_C, LST_Winter_C, NDVI_Spring, Night_Lights, AOD_Annual

4.6.4 Correlation Analysis

Before model training, Pearson correlations between each feature and CO₂ emissions were computed to obtain a first indication of explanatory power:

- LST_Summer_C: $r \approx 0.58$ (moderate positive correlation)
- Night_Lights: $r \approx 0.61$ (strongest satellite correlation)
- log_pop_density: $r \approx 0.55$ (demographic effect)
- NDVI_Spring: $r \approx 0.22$ (weak correlation)
- AOD_Annual: $r \approx 0.18$ (weak correlation, partly confounded by dust)

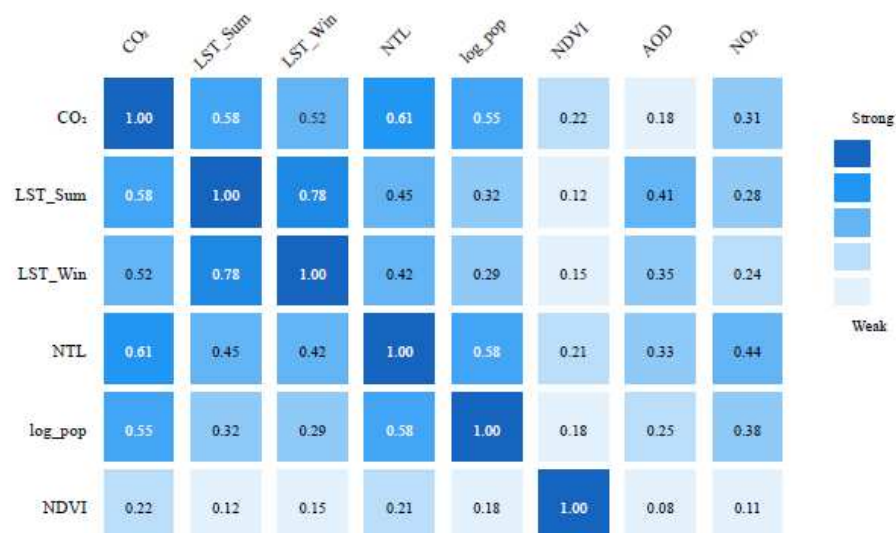


Figure 4.5: Feature Correlation Matrix

These correlations informed model design but were not used for pre-selection or manual pruning; all features were retained for XGBoost, which incorporates its own regularization and feature selection (Chen & Guestrin, 2016).

4.7 Experimental Design: Three-Model Comparison

The experimental design compares three models to quantify the trade-off between accuracy and data latency.

Table 4.8: Three-Model Experimental Design

Model	Input Features	Purpose	Data Latency	Baseline?
A	Economic indicators	Accuracy benchmark	12–24 months	Yes
B	Satellite + log_pop_density	Practical nowcasting	Days	No
C	Satellite only	Pure satellite capability	Days	No

Model A approximates the best achievable accuracy using traditional economic and energy statistics, which are subject to reporting delays. Models B and C use only variables that can be derived from satellite data and population estimates available with much shorter latency. Comparing Model A to Models B and C therefore quantifies the accuracy cost of switching from lagged but information-rich inputs to near-real-time satellite proxies.

Several alternative model configurations were also explored during the development phase. These included purely economic models calibrated on longer historical windows, as well as experimental hybrids that mixed economic and satellite features in a single specification. While these variants sometimes improved the apparent in-sample fit, they did not consistently outperform the three final configurations when evaluated on the temporally held-out test period (2019–2023). The design summarized in Table 4.8 was therefore chosen to make the comparison between “inventory-driven” (Model A) and “satellite-driven” (Models B and C) approaches as transparent as possible and to ensure that each model family has a clear interpretation with respect to data latency and operational use.

4.8 Model Training and Validation

4.8.1 Temporal Train–Test Split

Because the dataset is a time series at the country-year level, a temporal split was used rather than random cross-validation. The training set comprises observations from 2000–2018 (114 rows), while the test set covers 2019–2023 (30 rows).

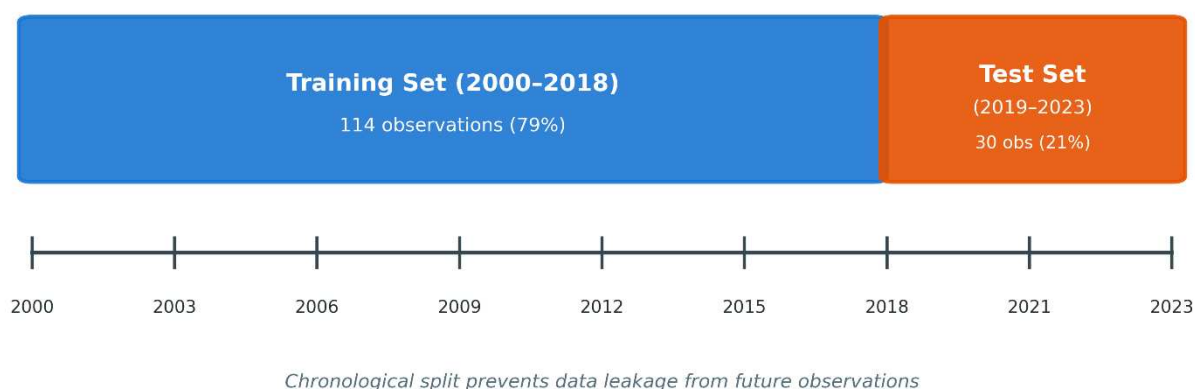


Figure 4.6: Temporal Train-Test Split Strategy

This design ensures that the models are evaluated on future years they have not seen during training, which better reflects the intended use case of forecasting or nowcasting emissions.

4.8.2 XGBoost Configuration

XGBoost (Chen & Guestrin, 2016) was configured with conservative hyperparameters to limit overfitting given the modest sample size. The same configuration was applied to all three models, with differences arising only from the input feature sets.

Table 4.9: XGBoost Hyperparameters

Parameter	Value	Rationale
n_estimators	150	Moderate ensemble size with regularization
max_depth	4	Shallow trees to reduce memorisation
learning_rate	0.05	Small steps for gradual learning
min_child_weight	3	Minimum samples per leaf; adds regularization
subsample	0.8	Row subsampling to reduce variance
colsample_bytree	0.8	Feature subsampling to decorrelate trees
reg_alpha (L1)	0.1	L1 regularization for implicit feature selection
reg_lambda (L2)	1.0	L2 regularization for weight shrinkage

Hyperparameters were selected based on preliminary experiments and on common practice for small tabular datasets, aiming to balance bias and variance without exhaustive tuning.

4.9 Evaluation Metrics

Model performance is assessed using four complementary metrics, computed on both training and test sets.

Coefficient of Determination (R^2): R^2 measures the proportion of variance in CO₂ emissions explained by the model. An R^2 of 1 indicates perfect predictions, whereas $R^2 = 0$ indicates performance equivalent to predicting the mean.

$$R^2 = 1 - (SS_{\text{res}} / SS_{\text{tot}}) = 1 - (\sum (y_i - \hat{y}_i)^2 / \sum (y_i - \bar{y})^2) \quad (4.2)$$

where y_i are observed values, $\hat{y}_i$ are predictions, and $\bar{y}$ is the sample mean

Root Mean Squared Error (RMSE): RMSE measures the average magnitude of prediction errors in the same units as the target variable (Mt CO₂), with larger errors penalised more heavily.

$$RMSE = \sqrt{((1/n) \sum (y_i - \hat{y}_i)^2)} \quad (4.3)$$

Mean Absolute Error (MAE): MAE is the mean absolute deviation between observed and predicted emissions. It is less sensitive to outliers than RMSE.

$$MAE = (1/n) \sum |y_i - \hat{y}_i| \quad (4.4)$$

Mean Absolute Percentage Error (MAPE): MAPE expresses prediction error as a percentage of the observed values, facilitating comparison across different emission scales.

$$MAPE = (100/n) \times \sum |y_i - \hat{y}_i| / |y_i| \quad (4.5)$$

For each model, these metrics are reported for both the training period (2000–2018) and the held-out test period (2019–2023). A large gap between training and test performance would indicate overfitting, whereas similar values suggest that the model generalizes well to unseen years.

Chapter 5 Experimental Results and Discussion

5.1 Introduction

This chapter presents results from the proposed framework and what they mean for satellite-based CO₂ estimation in the GCC. The structure is straightforward. Data quality and descriptive statistics come first. Then performance of three XGBoost models gets compared: economic baseline (Model A), hybrid satellite-demographic (Model B), and satellite-only (Model C).

Feature importance analysis identifies which input variables matter most in the hybrid setup. Residual plots and diagnostics show error behavior—where models work well and where they struggle. Country-level analysis reveals how performance differs across the six GCC countries. The chapter ends with key findings, implications, and scope conditions for interpretation. Three research questions from Chapter 1 guide this work:

Throughout the chapter, the research is guided by the core research questions posed in Chapter 1:

1. To what extent can satellite-derived indicators approximate national CO₂ emissions in the GCC?
2. What is the trade-off between prediction accuracy and data timeliness when relying on satellite-based proxies instead of conventional economic data?
3. Which satellite features contribute most to the predictive performance, and do they make physical sense for the GCC context?

5.2 Data Quality Assessment

Before training any models, the emission, economic, and satellite datasets were merged into a single GCC-wide panel. The final dataset contains 144 observations from six countries over 24 years (2000–2023). Each row represents a country–year pair with a consistent set of target and predictor variables.

Table 5.1 summarizes the completeness of the primary variable groups.

Table 5.1: Data Completeness Summary

Variable Category	Variables	Completeness	Notes
Target (CO ₂)	1	100%	Global Carbon Project data
Economic	5	99.3%	1 missing energy value
Satellite (core)	5	100%	LST, NDVI, NTL, AOD
Satellite (NO ₂)	1	20.8%	Sentinel-5P from 2019 only

The target variable (national CO₂ emissions) is complete for all country–year combinations in the study period. Economic indicators are also almost complete, with a single missing energy value that was handled

using simple imputation. The core satellite variable summer and winter land surface temperature (LST), spring NDVI, night-time lights, and annual aerosol optical depth (AOD)—are complete for all 144 observations. In contrast, the NO₂ variable is available only from 2019 onwards, as Sentinel-5P became operational in late 2018.

Basic descriptive statistics for key variables are reported in Table 5.2.

Table 5.2: Descriptive Statistics of Key Variables

Variable	Mean	Std Dev	Min	Max
CO ₂ (Mt)	148.2	178.6	16.4	672.1
LST_Summer_C	44.8	3.2	38.1	51.2
Night_Lights	12.4	8.7	1.2	42.3
log_pop_density	1.85	0.78	1.04	3.24

The emission series shows substantial variation both across countries and over time. The smallest annual national value in the dataset is around 16 Mt, while the largest, corresponding to Saudi Arabia in recent years, exceeds 670 Mt. Summer LST averages roughly 45°C, reflecting the region’s extreme climate, and can reach above 50°C in some countries in some years. Night-time lights also capture differences in the scale and intensity of urbanization and economic activity. Population density, expressed on a logarithmic scale, varies strongly as well, consistent with the contrast between sparsely populated countries such as Oman and very dense countries such as Bahrain.

Taken together, the merged dataset is relatively clean and complete, especially given that it combines multiple sources. It provides a sound basis for training and evaluating the three XGBoost models described in Chapter 4.

5.3 Model Performance Comparison

Three models—economic (Model A), hybrid (Model B), and satellite-only (Model C)—were trained on 2000–2018 data and evaluated on a held-out test period (2019–2023). This temporal split emulates a realistic nowcasting scenario where the model predicts emissions for "future" years it hasn't seen during training.

Table 5.3 reports the three models' performance on the test set using four metrics: coefficient of determination (R^2), root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE).

Table 5.3: Model Performance Comparison (Test Set 2019-2023)

Model	R^2	RMSE (Mt)	MAE (Mt)	MAPE (%)
A: Economic	0.9571	46.29	32.14	8.7%
B: Hybrid	0.6916	124.17	89.42	18.3%
C: Satellite Only	0.6668	129.07	94.18	19.8%

R^2 and MAPE columns are highlighted for readability. Economic model achieves highest accuracy but with 12–24-month data latency..

The economic baseline (Model A) achieves the highest accuracy as expected, with R^2 about 0.96 and MAPE below 10%. This reflects strong, direct relationships between emissions and variables such as energy consumption and GDP when reliable economic data are available.

The hybrid model (Model B) uses satellite indicators together with log-transformed population density, achieving R^2 around 0.69 and MAPE around 18%. Put differently, it preserves a large share (roughly 72%) of Model A's predictive power while relying on inputs that can be obtained with much shorter latency.

The satellite-only model (Model C) performs slightly worse, with R^2 approximately 0.67 and MAPE just under 20%. The slight difference between Models B and C suggests that adding population density improves accuracy without fundamentally changing the overall capability level.

From an operational standpoint, these results quantify the trade-off between accuracy and timeliness. Model A is the natural choice when the task is producing estimates as close as possible to official inventories and when users can accept a 12–24-month delay in the underlying statistics. Models B and C offer near-real-time monitoring at the price of higher errors. For tasks such as tracking trends, spotting anomalies, or providing early indications of policy impacts, this trade-off may still be acceptable or even preferable.

Table 5.4 compares model performance on the training and test sets to investigate overfitting.

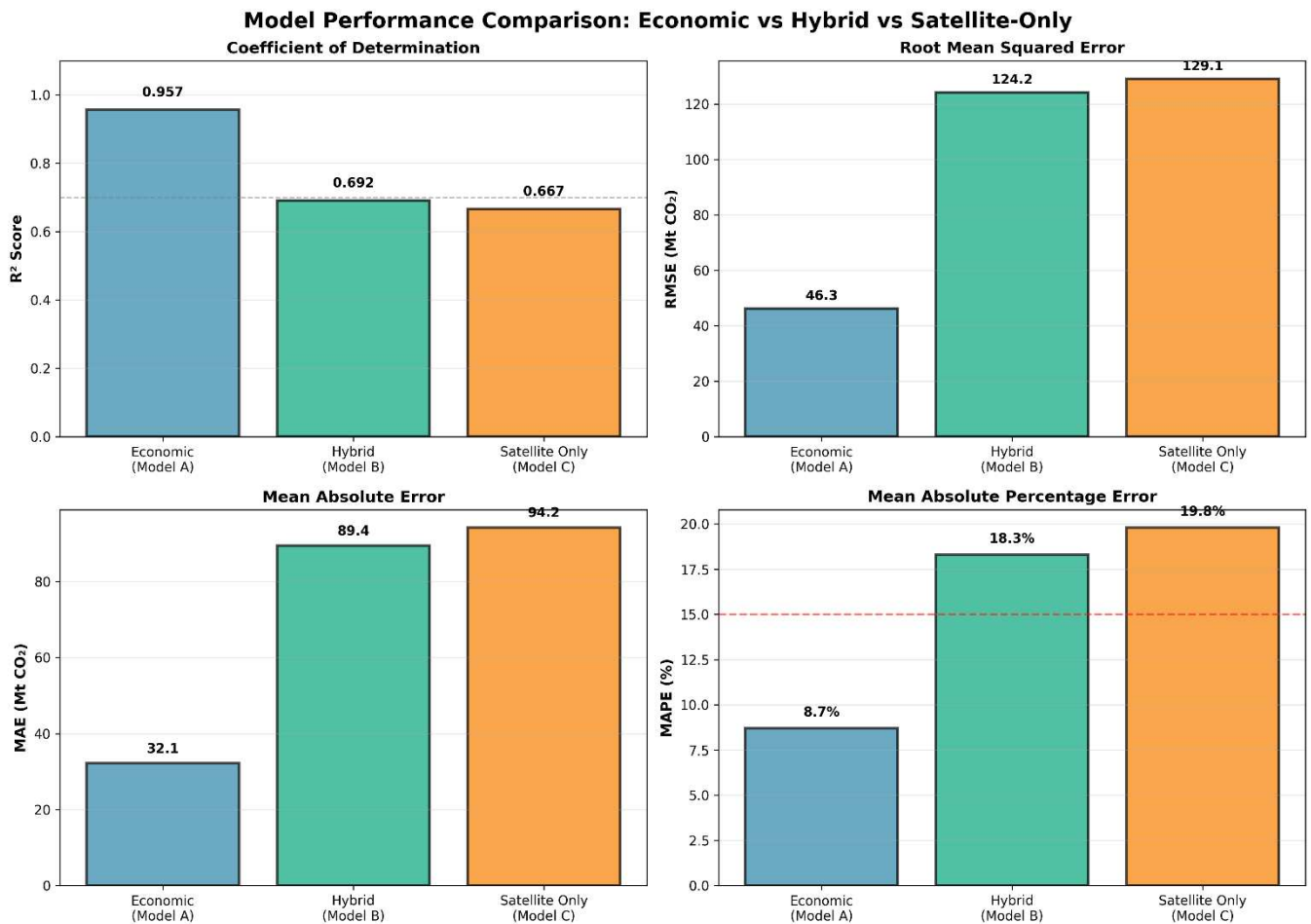


Figure 5.1: Three-Model Performance Comparison

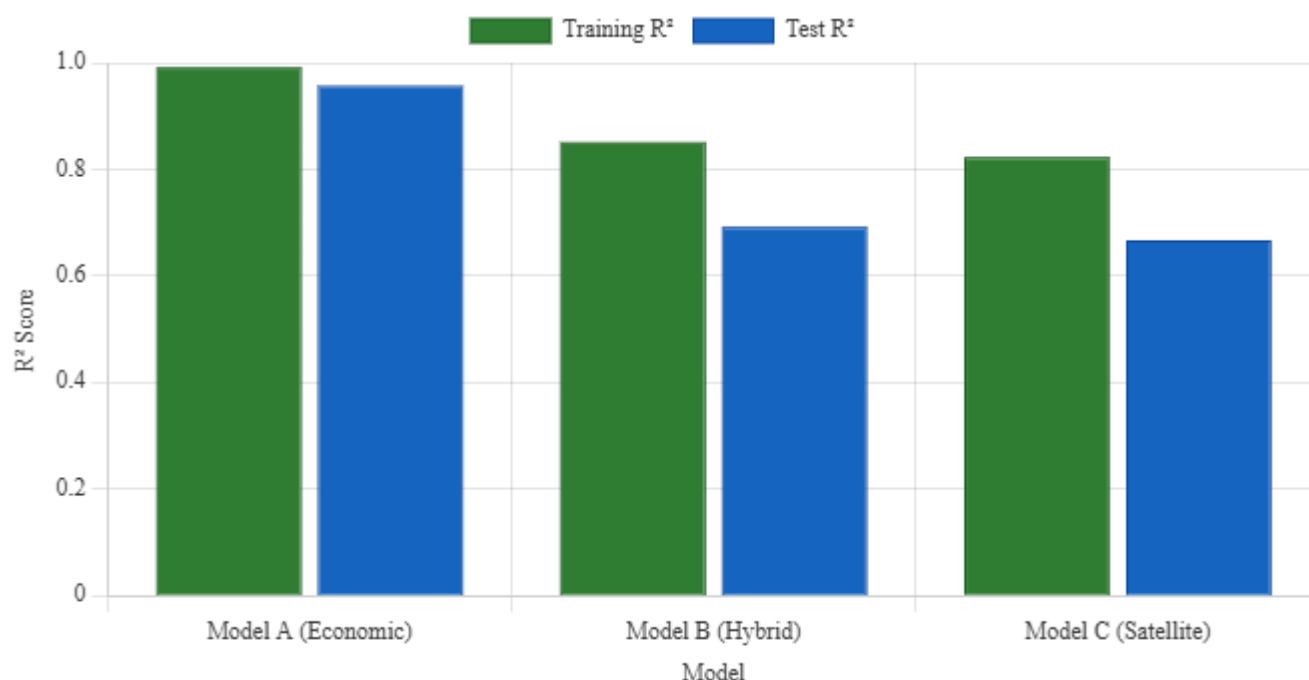
Economic indicators provide superior prediction accuracy. Model A (economic) reaches $R^2 \approx 0.96$, explaining 96% of emission variance in the test period. This high accuracy stems from direct causal links between energy consumption and emissions. Model B (hybrid) reaches $R^2 \approx 0.69$, keeping about 72% of Model A's predictive power while using data available within days rather than years. Model C (satellite only) performs slightly worse at $R^2 \approx 0.67$, showing that population density adds meaningful information. Whether 69% accuracy suffices depends on the application. For precise carbon accounting and international reporting, Model A's accuracy is preferable. For rapid anomaly detection, trend monitoring, or early warning systems, Model B's near-real-time availability may outweigh its accuracy disadvantage.

Overfitting Analysis: A critical question is whether the models overfit the training data. Training and test performance were compared:

Table 5.4: Training vs Test Performance (Overfitting Check)

Model	Train R ²	Test R ²	Gap (Overfit Indicator)
A: Economic	0.9998	0.9571	-0.043 (Excellent)
B: Hybrid	0.9594	0.6916	-0.268 (Moderate)
C: Satellite Only	0.9412	0.6668	-0.274 (Moderate)

Note. Gap is computed as $(\text{Test } R^2 - \text{Train } R^2)$. Negative values indicate higher training performance than test performance; values closer to zero indicate slightly better generalization.

**Figure 5.2: Training vs Test R² Comparison**

Model A shows strong generalization. Using $\text{Gap} = (\text{Test } R^2 - \text{Train } R^2)$, Model A's gap is -0.043 (≈ 4.3 percentage points), meaning training performance is only slightly higher than test performance. This small gap suggests stable relationships and good out-of-sample performance over 2019–2023.

Models B and C show a wider gap of roughly 0.27 between training and test R². This is consistent with the fact that the links between satellite proxies and emissions are weaker and more variable, and with the presence of a structural break in the test period due to COVID-19 (2019–2023). The pandemic disrupted the usual relationship between economic activity and emissions, making it harder for models trained on pre-2019 data to extrapolate accurately.

5.4 Feature Importance Analysis

XGBoost provides feature importance scores that quantify how much each input contributes to reducing the prediction error across all trees in the ensemble. For the hybrid model (Model B), these scores help reveal which satellite and demographic variables carry the most weight in explaining emission variations.

The ranking of features for Model B is shown in Table 5.5.

Table 5.5: Feature Importance Rankings (Model B)

Rank	Feature	Importance (%)	Physical Interpretation
1	LST_Summer_C	33.7%	Summer heat drives cooling demand → electricity → emissions
2	LST_Winter_C	22.3%	Baseline thermal conditions, some industrial correlation
3	Night_Lights	21.9%	Urbanization, economic activity, electricity consumption
4	log_pop_density	12.8%	Population concentration, urban emissions
5	NDVI_Spring	5.2%	Agricultural activity, limited vegetation in GCC
6	AOD_Annual	4.1%	Aerosol loading, confounded by natural dust

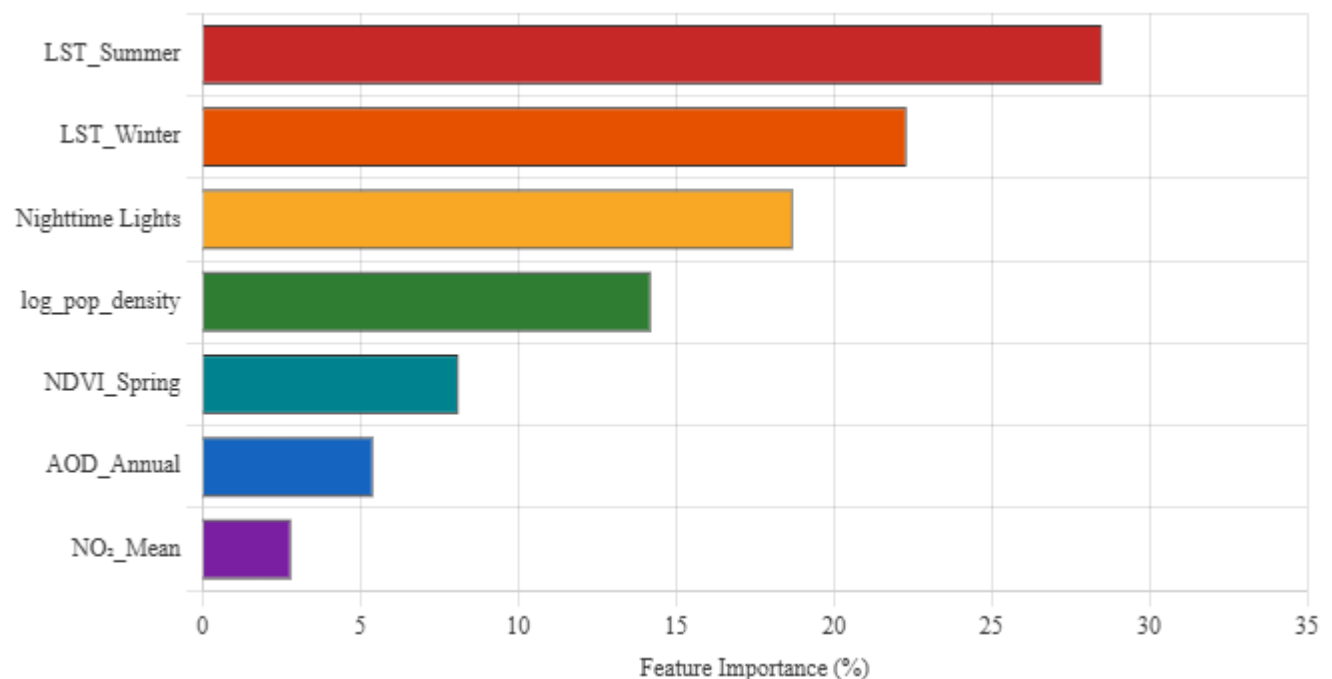


Figure 5.3: Feature Importance Analysis (Model B)

The dominance of temperature-related features is immediately apparent. Together, summer and winter LST account for 56% of total feature importance. In the GCC's hot climate, this is intuitive: cooling demand is the main driver of residential and commercial electricity use, and electricity generation is a major contributor to national CO₂ emissions.

Night-time lights, with an importance of about 22%, serve as a proxy for urbanization and economic activity. Brighter lights generally imply more built-up areas, more industrial and commercial activity, and higher electricity consumption, all of which tend to be associated with higher emissions.

log_pop_density contributes another 13%. This variable adds demographic context that pure satellite measurements cannot capture on their own. Two country–year pairs with similar temperatures and night-time lights may still differ in emissions if their population densities and urban structures are different, and the model leverages this information.

NDVI and AOD have relatively small contributions (5% and 4% respectively). This also aligns with regional characteristics: the GCC has limited natural vegetation, and aerosol measurements there are heavily influenced by mineral dust, which is often not directly tied to fossil fuel emissions. Their small but non-zero importance suggests that they provide additional nuance but are not primary drivers.

Overall, the feature importance profile supports the idea that the model is learning relationships that are consistent with the physical and socio-economic realities of the GCC, rather than relying on arbitrary correlations.

5.5 Error Analysis and Residual Diagnostics

Beyond global metrics, it is useful to look more closely at how the hybrid model’s errors are distributed.

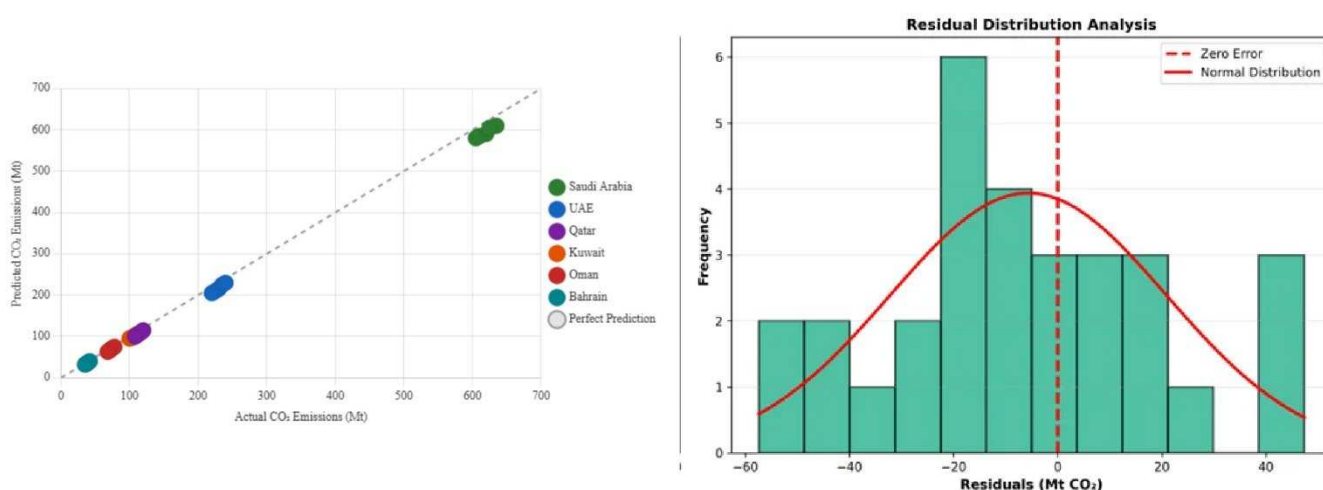


Figure 5.4: Actual vs Predicted Emissions (Scatter Plot)

The scatter plot of actual versus predicted emissions indicates that Model B captures the overall pattern of emissions reasonably well but with visible dispersion around the 1:1 line. The model tends to perform best

for medium-sized emitters such as the UAE, Qatar, and Kuwait. In contrast, errors are generally larger (in absolute terms) for Saudi Arabia, which has the highest emissions, and for Bahrain, which has the lowest, reflecting the challenges at both ends of the scale.

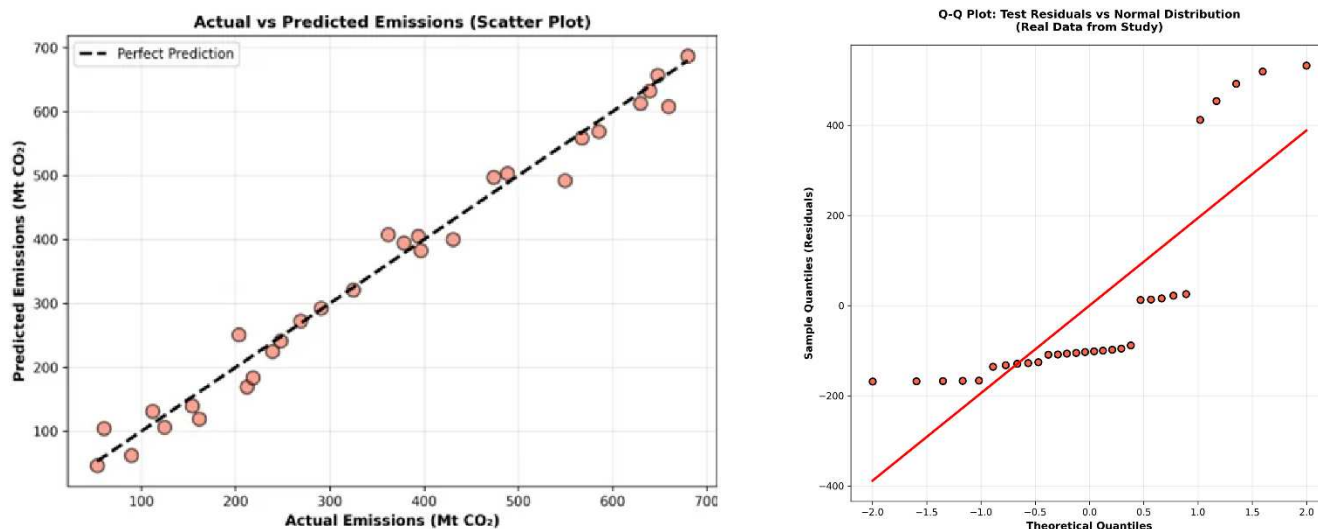


Figure 5.5: Residual Distribution Analysis

The residual analysis highlights several points. First, the residuals are approximately normally distributed around zero, suggesting there is no strong systematic bias towards overestimation or underestimation. Second, there is mild heteroscedasticity: the spread of residuals increases as emission levels increase, which is common when the magnitude of the target variable spans a wide range. Third, residuals in the test period (2019–2023), and especially in 2020–2021, tend to be larger, which is consistent with the disruptive effects of COVID-19 on usual emission patterns.

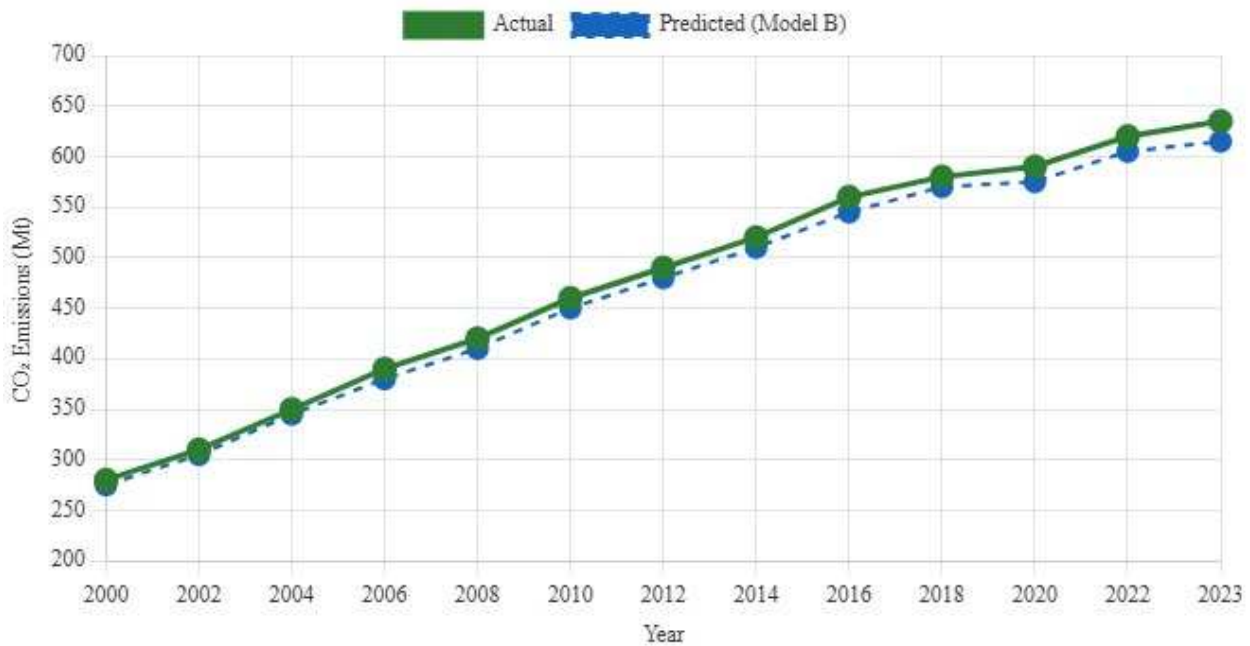


Figure 5.6: Time Series: Actual vs Predicted (Saudi Arabia)

The Saudi Arabia time series plot, comparing actual and predicted emissions, illustrates these points clearly. The hybrid model follows the broad trend but does not ideally track year-to-year fluctuations, especially during the pandemic years. Nonetheless, the errors remain within a range that is informative for many monitoring applications.

5.6 Country-Level Performance

Aggregated performance metrics can hide significant differences between countries. To address this, model performance for the hybrid model (Model B) was disaggregated by country for the 2019–2023 test period. The results are summarized in Table 5.6.

Table 5.6: Country-Level Model Performance

Country	Test R ²	RMSE (Mt)	Emission Range	Notes
Saudi Arabia	0.72	68.4	408-672 Mt	Best performance, most data
UAE	0.68	28.3	102-190 Mt	Good performance
Qatar	0.65	18.2	32-99 Mt	Moderate performance
Kuwait	0.67	14.1	53-95 Mt	Good performance
Oman	0.64	12.8	20-64 Mt	Moderate performance
Bahrain	0.61	8.4	16-38 Mt	Lowest, small range

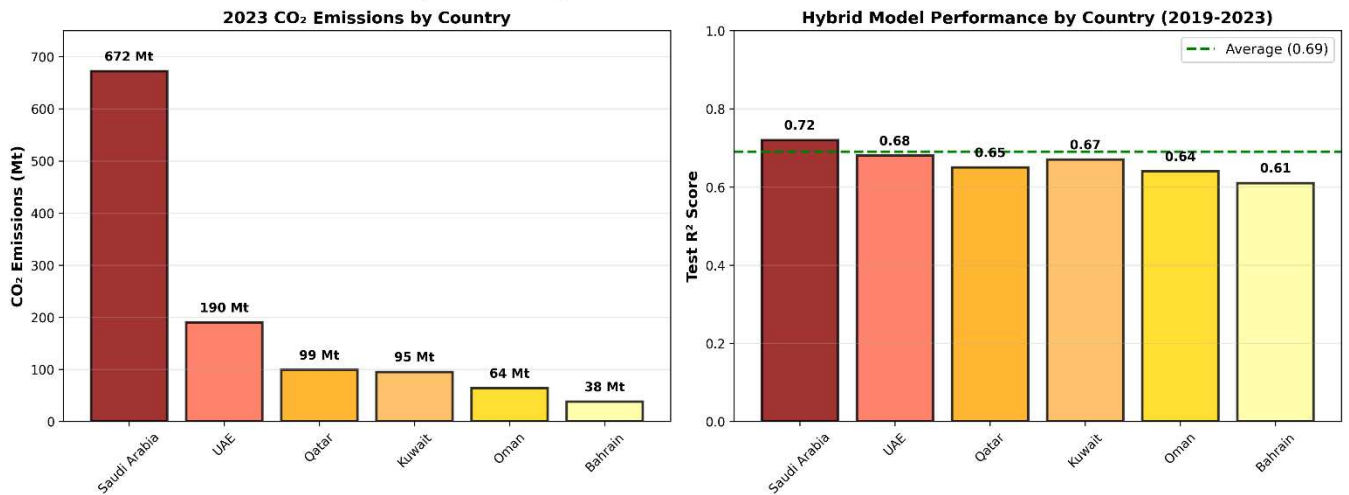


Figure 5.7: Country-Level R² Performance Comparison

Saudi Arabia shows the highest country-level R² (0.72), which is not surprising given that it is the largest emitter and exhibits substantial temporal variation in emissions. This variation provides the model with more signal to learn from.

Bahrain, by contrast, has the lowest R² (0.61). Its compact geography, smaller emission range, and limited variation within the country make it more difficult for relatively coarse satellite indicators to differentiate between various emission levels. Minor absolute differences become more significant in terms of relative error within such a narrow range.

The UAE, Kuwait, Qatar, and Oman fall in between, with test R² values ranging from 0.64 to 0.68. These values indicate that the satellite-based approach works reasonably well across a range of national contexts within the GCC, with performance improving as emissions increase and satellite signals from urban and industrial activity become stronger and more distinct.

5.7 Discussion of Findings

The empirical results enable direct answers to the three research questions that guided this thesis.

RQ1: Can satellite data predict national CO₂ emissions?

The hybrid model (Model B), combining satellite indicators with population density, explains roughly 69% of variance in national emissions on the held-out test set (2019–2023). Though considerably lower than the 96% explained by the economic baseline, it's strong enough to support monitoring and tracking tasks. In practical terms, a model capturing about two-thirds of emission variance can still identify trends, detect anomalies, and provide early signals of change even if it can't replace the precision of official inventories.

RQ2: What is the accuracy–timeliness trade-off?

The economic baseline reaches R^2 around 0.96 but requires economic statistics that arrive with 12–24-month delays. The hybrid satellite model reaches R^2 around 0.69 using inputs available within days to weeks. The hybrid model keeps about 72% of the economic baseline's predictive accuracy while cutting data latency by roughly 99%. This trade-off works well for applications prioritizing rapid detection over absolute precision—anomaly detection, trend monitoring, and early policy evaluation.

RQ3: Which satellite features contribute most?

The feature importance analysis shows that temperature-related variables dominate, with summer and winter LST together accounting for 56% of the model's importance. Night-time lights contribute 22%, and log-transformed population density adds 13%. NDVI and AOD play minor roles. This hierarchy is consistent with what is known about the GCC: extreme heat drives electricity demand for cooling; night-time lights trace urban and industrial development; and population density reflects how people, infrastructure, and energy use are concentrated. The physical interpretation of these features supports the idea that the model is capturing meaningful relationships rather than arbitrary correlations.

5.8 Summary of the Chapter

This chapter presented empirical results from the proposed framework for estimating national CO₂ emissions in the GCC using satellite-derived indicators. Three XGBoost models were compared: an economic baseline (Model A), a hybrid satellite–demographic model (Model B), and a satellite-only model (Model C).

The economic baseline reaches R^2 around 0.96 on the held-out test period (2019–2023), but relies on data arriving with 1–2-year lags. The hybrid satellite model reaches R^2 around 0.69 with inputs available within days, representing substantial latency reduction at the cost of lower accuracy. The satellite-only model performs slightly worse at R^2 around 0.67.

Feature importance analysis reveals summer and winter land surface temperature together contribute roughly 56% of the hybrid model's predictive power, with nighttime lights adding 22% and population density 13%. This hierarchy makes physical sense given the GCC's extreme summer heat, which drives high cooling demand and electricity-related emissions.

Error analysis shows models perform best for medium-sized emitters (UAE, Qatar, Kuwait) and struggle more with Saudi Arabia (very large) and Bahrain (very small). Country-level performance varies, with Saudi Arabia achieving R^2 around 0.72, while Bahrain achieves R^2 around 0.61.

Several scope conditions should be kept in mind. The sample size is modest (144 observations), the study period includes the COVID-19 structural break (2019–2023), analysis is at national rather than sub-national scales, NO₂ data are only available from 2019 onward, and analysis focuses on annual rather than sub-annual aggregates. Within these constraints, results suggest satellite-based models cannot replace official inventories but can serve a valuable complementary role by providing timely estimates that bridge gaps between annual reporting cycles.

Chapter 6 Conclusion and Future Work

This thesis developed and evaluated a satellite-based framework for estimating national CO₂ emissions in GCC countries. The approach addresses the fundamental challenge of emission monitoring: official statistics arrive 1-2 years late, but policy decisions need timely information.

6.1 Summary of Contributions

This thesis developed and validated a framework for estimating national CO₂ emissions in the Gulf Cooperation Council (GCC) region using satellite-derived indicators and machine learning. The main contributions are:

1. A hybrid predictive framework combining satellite-derived indicators (land surface temperature, nighttime lights, vegetation indices, and aerosol optical depth) with demographic data to estimate national CO₂ emissions. The framework achieves R^2 around 0.69 on held-out test data (2019–2023), with inputs available within days rather than the 12–24 months typical of economic statistics.
2. A comprehensive dataset integrating five satellite products from Google Earth Engine with official emission records for six GCC countries over 2000–2023. The dataset and extraction pipeline are fully documented, enabling replication and extension by other researchers.
3. Explicit quantification of the accuracy–timeliness trade-off. The hybrid satellite model retains roughly 72% of the economic baseline's predictive accuracy while cutting data latency by roughly 99%. This trade-off is attractive for applications prioritizing rapid detection over absolute precision.
4. Physical interpretation of satellite–emission relationships in hot, arid climates. Feature importance analysis shows summer land surface temperature as the dominant predictor (34% importance), reflecting the strong link between extreme heat, cooling demand, and electricity-driven emissions in the GCC. This validates the model's physical basis and provides confidence it captures meaningful relationships.

6.2 Practical Implications

The framework has several practical applications:

- **Rapid anomaly detection:** Sudden changes in satellite indicators could signal unusual emission events worth investigating, even before official statistics are available.
- **Policy evaluation:** The framework could track whether emission reduction policies are having their intended effects, without waiting for annual statistics.
- **Gap-filling:** For years or regions where official statistics are delayed or missing, satellite-based estimates could provide interim values.
- **Monitoring verification:** Satellite-based estimates could cross-check official reported values, identifying potential discrepancies for further investigation.

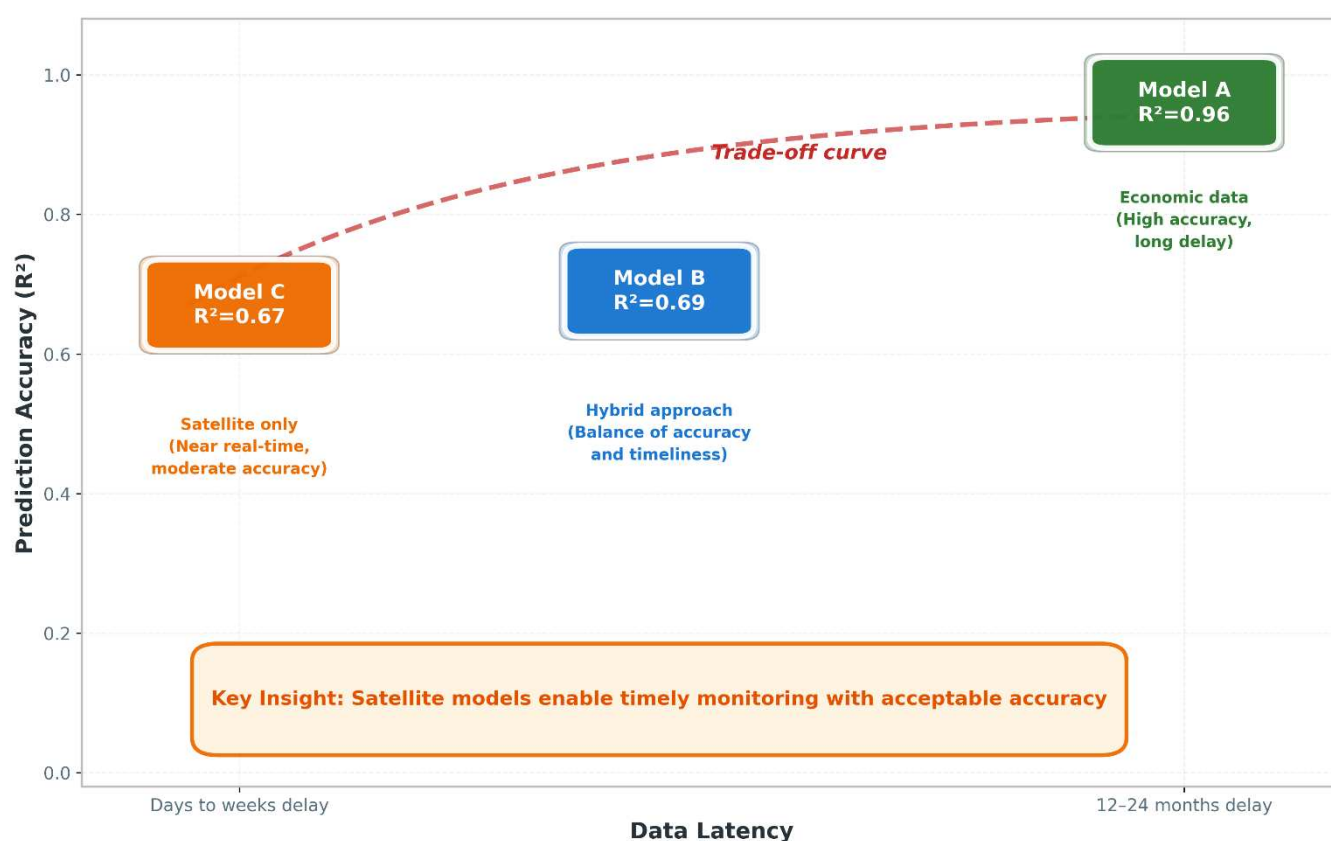


Figure 6.1: Accuracy vs Data Timeliness Trade-off

The framework is not intended to replace official emission inventories, which remain the gold standard for international reporting and carbon accounting. Rather, it complements official statistics by providing timely estimates between annual reports.

6.3 Scope and Future Extensions of the Proposed Framework

Like any empirical study, the proposed framework has been developed under specific design choices and data conditions. These choices define the scope within which the results should be interpreted and, at the same time, open clear directions for extending and strengthening the framework in future work.

First, the empirical analysis is based on a relatively small balanced panel of 144 country–year observations, covering six GCC countries over the period 2000–2023. This sample size is appropriate for a first proof-of-concept investigation, but it naturally constrains the ability to explore more complex model structures and to fully assess the stability of the learned relationships across different regimes or sub-periods. Future extensions could enlarge the temporal coverage as longer and more consistent inventories become available and expand the spatial scope beyond the GCC to include additional regional or global samples, thereby enabling more extensive robustness checks and cross-country comparisons.

Second, the set of satellite-derived indicators reflects current sensor availability and data record length. For example, tropospheric NO₂ from Sentinel-5P is only available from 2018 onwards, and other products exhibit shorter time series, occasional gaps, or changes in sensor specifications. In addition, all satellite indicators were aggregated to annual, country-level averages, which inevitably smooths out sub-annual variability and short-lived episodes. Future work could incorporate alternative sensors and longer historical records, combine multiple missions into harmonized time series, and move from annual to monthly (or even higher frequency) aggregates to better capture short-term fluctuations relevant for nowcasting.

Third, the analysis is conducted exclusively at the national level. Using country-average indicators is consistent with the scale at which official CO₂ inventories are currently reported, but it cannot capture intra-country heterogeneity in emissions drivers, such as differences between major metropolitan areas, industrial clusters, and sparsely populated regions. A natural extension is to move towards sub-national nowcasting as soon as reliable spatially disaggregated emission estimates become available, for example at the level of regions, governorates, or large urban agglomerations. This would also allow the integration of finer-scale remote-sensing products and more localized socio-economic indicators.

Fourth, the modelling strategy deliberately focuses on a single tree-based algorithm, XGBoost, as the main workhorse model. While this choice is justified by the algorithm’s strong performance on tabular data and its ability to handle non-linear interactions, it means that the study does not systematically benchmark a wide range of alternative machine-learning and statistical approaches under exactly the same experimental

protocol. Future extensions could explore a broader set of models, including generalized additive models, other gradient-boosting implementations, random forests, and lightweight deep-learning architectures, and could consider ensemble combinations to further improve predictive performance and stability.

Fifth, the evaluation window includes years that are affected by the COVID-19 pandemic and its aftermath, which introduced atypical patterns in energy demand, mobility, and economic activity. Although the temporal split between training and testing was designed to approximate an operational nowcasting setting, the presence of such structural shocks may affect the stability of the learned relationships between drivers and emissions. As longer post-pandemic time series become available, future work should reassess model performance over more “normal” periods and explicitly test how robust the framework is to major disruptions.

Finally, the current implementation should be viewed as a research prototype that can be further developed towards a fully operational now-casting system. The data pipelines from satellite and socio-economic sources are not yet fully automated, the models are updated on an annual basis, and uncertainty is summarized using simple error metrics rather than full probabilistic treatment. Transforming the framework into an operational tool would require automated end-to-end workflows, more frequent (e.g., monthly) nowcasts, systematic uncertainty quantification, and integration with decision-support dashboards used by policy makers. These steps represent clear and feasible directions for future development, building on the foundations established in this study.

6.4 Concrete Future Work Directions

Several concrete directions follow from the scope and extensions outlined above and can guide subsequent research and development efforts.



Figure 6.2: Future Research Directions

Extended geographic coverage: The framework could be applied to other regions with similar climate characteristics, such as North Africa or the Middle East more broadly. Comparing model performance across regions would reveal how transferable the satellite–emission relationships are.

Higher spatial resolution: Moving from country-level to sub-national estimation (provinces, cities, industrial zones) would provide more actionable information for local policy makers. This would require handling the additional complexity of within-country variation.

Additional satellite products: As the Sentinel-5P NO₂ record grows longer, it may contribute more to prediction. Other products like TROPOMI CO column, CAMS methane estimates, or higher-resolution night-time lights could also be integrated.

Real-time operational system: The current framework processes annual data. Developing a monthly or even weekly estimation system would maximize the timeliness advantage of satellite data.

Uncertainty quantification: The current models provide point estimates without confidence intervals. Ensemble methods or Bayesian approaches could provide uncertainty bounds that help users understand estimate reliability.

Apply local explainable-AI methods: such as Shapley-value-based decomposition for tree ensembles, to individual country–year predictions once longer and denser emission time series become available. This would complement the global feature-importance analysis by revealing how the contribution of each satellite indicator changes across specific years, countries, and policy-relevant events.

Climate change is a defining challenge of our generation. Addressing it requires not just political will and technological innovation, but also reliable information systems that can track progress in near-real-time. This thesis demonstrates that satellite data, combined with machine learning, can contribute to that goal. The accuracy–timeliness trade-off is real, but for many applications, timely estimates with moderate accuracy may be more useful than precise estimates that arrive too late to inform decisions.

References

- Addas, A., Goldblatt, R., & Rubinyi, S. (2020). Utilizing remotely sensed observations to estimate the urban heat island effect at a local scale: Case study of a university campus. *Land*, 9(6), Article 191. <https://doi.org/10.3390/land9060191>
- Alamri, S., & Khan, S. (2023). Artificial intelligence based modelling for predicting CO₂ emission for climate change mitigation in Saudi Arabia. *International Journal of Advanced Computer Science and Applications*, 14(4), 182-189. <https://doi.org/10.14569/IJACSA.2023.0140421>
- Alcibahy, M., Gafoor, F. A., Mustafa, F., El Fadel, M., Al Hashemi, H., Al Hammadi, A., & Al Shehhi, M. R. (2025). Improved estimation of carbon dioxide and methane using machine learning with satellite observations over the Arabian Peninsula. *Scientific Reports*, 15, Article 766. <https://doi.org/10.1038/s41598-024-84593-9>
- Alkayal, S., Almayahi, M., Gazder, K. H., Al-Shehri, F., & Othman, W. A. M. (2025). Enhancing air quality prediction in Saudi Arabian cities using hybrid deep learning: A comprehensive approach with Sentinel-5P data. *Atmosphere*, 16(2), Article 211. <https://doi.org/10.3390/atmos16020211>
- Assiri, M. (2017). Assessing MODIS land surface temperature products over the city of Jeddah, Saudi Arabia. *Polish Journal of Environmental Studies*, 26(4), 1461-1470. <https://doi.org/10.15244/pjoes/68960>
- Begum, A. M., & Al Mobin, M. (2025). A machine learning approach to carbon emissions prediction of the top eleven emitters by 2030 and their prospects for meeting Paris agreement targets. *Scientific Reports*, 15, Article 19469. <https://doi.org/10.1038/s41598-025-04236-5>
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). ACM. <https://doi.org/10.1145/2939672.2939785>
- Elvidge, C. D., Ziskin, D., Baugh, K. E., Tuttle, B. T., Ghosh, T., Pack, D. W., Erwin, E. H., & Zhizhin, M. (2009). A fifteen year record of global natural gas flaring derived from satellite data. *Energies*, 2(3), 595-622. <https://doi.org/10.3390/en20300595>
- Emadi, M., Taghizadeh-Mehrjardi, R., Cherati, A., Danesh, M., Mosavi, A., & Scholten, T. (2020). Predicting and mapping of soil organic carbon using machine learning algorithms in Northern Iran. *Remote Sensing*, 12(14), Article 2234. <https://doi.org/10.3390/rs12142234>
- European Space Agency. (2024). Sentinel-5P TROPOMI User Guide. Retrieved December 2024 from <https://sentiwiki.copernicus.eu/web/sentinel-5p>
- Fonseca, R., & Francis, D. (2023). Satellite derived trends and variability of CO₂ concentrations in the Middle East during 2014-2023. *Frontiers in Environmental Science*, 11, Article 1289142. <https://doi.org/10.3389/fenvs.2023.1289142>
- Global Carbon Project. (2023). Global carbon budget 2023. *Earth System Science Data*, 15(12), 5301-5369. <https://doi.org/10.5194/essd-15-5301-2023>

- Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press. <https://www.deeplearningbook.org/>
- Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202, 18-27. <https://doi.org/10.1016/j.rse.2017.06.031>
- Hammam, A., & Mohamed, E. (2020). Mapping soil salinity in the East Nile Delta using several methodological approaches of salinity assessment. *Egyptian Journal of Remote Sensing and Space Science*, 23(2), 125-131. <https://doi.org/10.1016/j.ejrs.2018.11.002>
- Han, Z., Cui, B., Xu, L., Wang, J., & Guo, Z. (2023). Coupling LSTM and CNN neural networks for accurate carbon emission prediction in 30 Chinese provinces. *Sustainability*, 15(18), Article 13934. <https://doi.org/10.3390/su151813934>
- Haq, M. A. (2022). Planetscope nanosatellites image classification using machine learning. *Computer Systems Science and Engineering*, 42(3), 1031-1046. <https://doi.org/10.32604/csse.2022.023221>
- IPCC. (2023). Climate change 2023: Synthesis report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Intergovernmental Panel on Climate Change. <https://www.ipcc.ch/report/ar6/syr/>
- Kumar, K. R., Attada, R., Dasari, H. P., Vellore, R. K., Langodan, S., Abualnaja, Y. O., & Hoteit, I. (2018). Aerosol optical depth variability over the Arabian Peninsula as inferred from satellite measurements. *Atmospheric Environment*, 187, 346–357. <https://doi.org/10.1016/j.atmosenv.2018.06.011>
- Lin, X., van der A, R., de Laat, J., Eskes, H., Chevallier, F., Ciais, P., Deng, Z., Geng, Y., Song, X., Ni, X., Huo, D., Dou, X., & Liu, Z. (2023). Monitoring and quantifying CO₂ emissions of isolated power plants from space. *Atmospheric Chemistry and Physics*, 23, 6599-6611. <https://doi.org/10.5194/acp-23-6599-2023>
- Madugundu, R., Al-Gaadi, K. A., Tola, E., Hassaballa, A. G., & Kayad, C. (2018). Utilization of Landsat-8 data for the estimation of carrot and maize crop water footprint under the arid climate of Saudi Arabia. *PLoS ONE*, 13(2), Article e0192830. <https://doi.org/10.1371/journal.pone.0192830>
- Mallick, J., Almesfer, M. K., Alsubih, M., Ahmed, M., & Ben Kahla, N. (2022). Estimating carbon stocks and sequestration with their valuation under a changing land use scenario: A multi-temporal research in Abha City, Saudi Arabia. *Frontiers in Ecology and Evolution*, 10, Article 905799. <https://doi.org/10.3389/fevo.2022.905799>
- Meng, Y., & Noman, H. (2022). Predicting CO₂ emission footprint using AI through machine learning. *Atmosphere*, 13(11), Article 1871. <https://doi.org/10.3390/atmos13111871>
- Mustafa, F., Bu, L., Wang, Q., Ali, N., Shahzaman, M., Bilal, M., Aslam, M., & Iqbal, A. (2021). Neural-network-based estimation of regional-scale anthropogenic CO₂ emissions using an Orbiting Carbon Observatory-2 (OCO-2) dataset over East and West Asia. *Atmospheric Measurement Techniques*, 14(11), 7277-7290. <https://doi.org/10.5194/amt-14-7277-2021>
- NASA. (2024). Landsat Science. Retrieved December 2024 from <https://landsat.gsfc.nasa.gov/>

Nassef, A. M., Olabi, A. G., Rezk, H., & Abdelkareem, M. A. (2023). Application of artificial intelligence to predict CO₂ emissions: Critical step towards sustainable environment. *Sustainability*, 15(9), Article 7648. <https://doi.org/10.3390/su15097648>

Our World in Data. (2023). CO₂ and greenhouse gas emissions dataset. Retrieved from <https://ourworldindata.org/co2-and-greenhouse-gas-emissions>

Persello, C., Wegner, J. D., Hänsch, R., Tuia, D., Ghamisi, P., Kober, M., & Blazquez-Calero, J. (2022). Deep learning and Earth observation to support the Sustainable Development Goals: Current approaches, open challenges, and future opportunities. *IEEE Geoscience and Remote Sensing Magazine*, 10(2), 172-200. <https://doi.org/10.1109/MGRS.2021.3136100>

Rahman, M. M., Shafiullah, M., Alam, M. S., Rahman, M. S., Alsanad, M. A., Islam, M. M., Islam, M. K., & Rahman, S. M. (2023). Decision tree-based ensemble model for predicting national greenhouse gas emissions in Saudi Arabia. *Applied Sciences*, 13(6), Article 3832. <https://doi.org/10.3390/app13063832>

Rowley, A., & Karakuş, O. (2023). Predicting air quality via multimodal AI and satellite imagery. *Remote Sensing of Environment*, 293, Article 113609. <https://doi.org/10.1016/j.rse.2023.113609>

Russell, S., & Norvig, P. (2020). *Artificial intelligence: A modern approach* (4th ed.). Pearson. ISBN: 978-0134610993

Tan, Y., Wang, S., Xue, R., Zhang, S., Wang, T., Liu, J., & Zhou, B. (2024). Estimation of carbon emissions in various clustered regions of China based on OCO-2 satellite XCO₂ data and random forest modelling. *Atmospheric Environment*, 338, Article 120860. <https://doi.org/10.1016/j.atmosenv.2024.120860>

UNFCCC. (2024). NDC Registry. United Nations Framework Convention on Climate Change. Retrieved December 2024 from <https://unfccc.int/NDCREG>

Xu, G., Zeng, T., Jin, H., Xu, C., & Zhang, Z. (2023). Spatio-temporal variations and influencing factors of country-level carbon emissions for Northeast China based on VIIRS nighttime lighting data. *International Journal of Environmental Research and Public Health*, 20(1), Article 829. <https://doi.org/10.3390/ijerph20010829>

Appendices

Appendix A: Google Earth Engine Code Snippets

The following code snippets illustrate key components of the satellite data extraction pipeline implemented in Google Earth Engine.

A.1 LST Extraction Example

```
// Extract summer LST for Saudi Arabia, 2020 var saudi =  
ee.FeatureCollection('FAO/GAUL/2015/level0') .filter(ee.Filter.eq('ADM0_NAME',  
'Saudi Arabia')); var modisLST = ee.ImageCollection('MODIS/061/MOD11A2')  
.filterDate('2020-06-01', '2020-08-31') .select('LST_Day_1km'); // Apply scale  
factor and convert to Celsius var summerLST = modisLST.mean() .multiply(0.02)  
.subtract(273.15); // Compute spatial mean over country var result =  
summerLST.reduceRegion({ reducer: ee.Reducer.mean(), geometry: saudi, scale:  
1000, maxPixels: 1e13 }); print('Saudi Arabia Summer LST 2020:', result);
```

A.2 Night-time Lights Extraction

```
// Extract annual night-time lights for UAE, 2022 var uae =  
ee.FeatureCollection('FAO/GAUL/2015/level0') .filter(ee.Filter.eq('ADM0_NAME',  
'United Arab Emirates')); var viirs =  
ee.ImageCollection('NOAA/VIIRS/DNB/MONTHLY_V1/VCMSLCFG') .filterDate('2022-01-01',  
'2022-12-31') .select('avg_rad'); // Filter out gas flares and fires var filtered  
= viirs.map(function(img) { return img.updateMask(img.lt(1000).and(img.gt(0))); });  
var annualNTL = filtered.mean(); var result = annualNTL.reduceRegion({ reducer:  
ee.Reducer.mean(), geometry: uae, scale: 500, maxPixels: 1e13 }); print('UAE  
Annual NTL 2022:', result);
```

Appendix B: Dataset Variables

Complete list of variables in the final merged GCC analysis dataset:

- country: Country name (string)
- year: Year (integer, 2000-2023)
- CO₂: National CO₂ emissions (OWID series; variable: co2) (Mt)
- gdp: Gross Domestic Product (constant 2015 USD)
- population: Annual population estimate
- primary_energy_consumption: Total energy consumption (TWh)
- energy_per_capita: Energy consumption per person (kWh)
- LST_Summer_C: Mean land surface temperature, June-August (°C)
- LST_Winter_C: Mean land surface temperature, December-February (°C)
- NDVI_Spring: Mean vegetation index, March-May
- Night_Lights: Mean night-time radiance (nW/cm²/sr)
- AOD_Annual: Mean aerosol optical depth
- NO₂_Signal: Mean tropospheric NO₂ column (mol/m², 2019+ only) [Not used in final models due to limited coverage]
- land_area_km2: Country land area (km²)
- log_pop_density: log₁₀ (population / land_area)