Analyzing Greenhouse Gas Emissions: Trends Predictions and Patterns

A CASE STUDY REPORT

Submitted by

Diva Alspeshkumar Merja (RA2211003011034)
Yaksh Vijaykumar Makadia (RA2211003011035)
Krishna Wadhwani (RA2211003011045)

For the course

Data Science - 21CSS303T

In partial fulfillment of the requirements for the degree of

BACHELOR OF TECHNOLOGY

DEPARTMENT OF COMPUTING TECHNOLOGIES

SCHOOL OF COMPUTING

FACULTY OF ENGINEERING AND TECHNOLOGY

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

KATTANKULATHUR - 603 203.

MAY 2025

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY
KATTANKULATHUR – 603 203

BONAFIDE CERTIFICATE

Certified that Data Science [21CSS303T], A Case Study Report titled "**Analysing global greenhouse gas emissions**" is the bonafide work of Diva Alspeshkumar Merja(RA2211003011034), Yaksh Vijaykumar Makadia (RA2211003011035), Krishna Wadhwani (RA2211003011045),who carried out the case study under my supervision. Certified further, that to the best of my knowledge the work reported herein does not form any other work.

Faculty Signature

Dr. S. Padmini
Associate Professor
Department of Computing Technologies.

Date: 1-05-2025

ABSTRACT

Climate change remains one of the most pressing and complex challenges of our time, with greenhouse gas (GHG) emissions being a primary driver of global warming and its cascading effects on ecosystems, weather patterns, and human livelihoods. As the world's third-largest emitter of GHGs, India faces a unique and formidable challenge in balancing its rapid economic growth with the urgent need to reduce emissions across key sectors such as energy, agriculture, industry, and transportation. This project seeks to address this critical issue by leveraging advanced machine learning techniques, specifically a Dense Neural Network (DNN) and Long Short-Term Memory (LSTM) models, to predict sector-wise GHG emissions for India based on comprehensive historical emission data.

By analyzing long-term trends, identifying key drivers of emissions, and forecasting future trajectories, the study aims to provide actionable insights that can inform evidence-based policy-making and targeted intervention strategies, all of which are essential for achieving the United Nations Sustainable Development Goals (SDG 13: Climate Action). The project employs a multi-faceted approach, combining data visualization to reveal patterns and anomalies, trend analysis to understand historical behavior, and predictive modeling to project emissions up to the year 2050 under various scenarios. Through this rigorous analysis, the study highlights which sectors—such as energy production, industrial activity, or agricultural practices—require immediate focus and mitigation efforts to align India's development trajectory with global climate targets. Additionally, the project explores the potential impact of policy interventions, technological advancements, and behavioral changes on emission reduction, offering a holistic view of how India can transition toward a low-carbon economy without compromising its developmental aspirations.

The findings of this research are intended to serve as a valuable resource for policymakers, environmentalists, and stakeholders, enabling them to prioritize actions, allocate resources effectively, and monitor progress toward sustainable development and climate resilience in the decades to come.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

GHG: **Greenhouse Gas**
LSTM: **Long Short-Term Memory Network**
DNN: **Dense Neural Network**
SDG: **Sustainable Development Goal**
CO2: **Carbon Dioxide**
CH4: **Methane**
N2O: **Nitrous Oxide**
RMSE: **Root Mean Squared Error**
MSE: **Mean Squared Error**

CHAPTER 1

1. INTRODUCTION

1.1 General

Climate change has emerged as one of the most critical threats to humanity, biodiversity, and ecosystems. A major contributor to climate change is the emission of greenhouse gases (GHGs), including carbon dioxide (CO_2), methane (CH_4), and nitrous oxide (N_2O). These gases trap heat within the Earth's atmosphere, leading to global warming and related adverse environmental effects such as rising sea levels, extreme weather events, and biodiversity loss.

India, with its rapid industrialization, urbanization, and population growth, ranks as the third-largest emitter of GHGs globally. Balancing economic development with environmental sustainability presents a significant challenge for policymakers. Sectors like energy production, transportation, agriculture, and industrial processes contribute substantially to India's GHG emissions.

Predicting GHG emissions with a high degree of accuracy is crucial to framing effective climate mitigation strategies. Reliable forecasts allow governments, industries, and international organizations to make informed decisions about resource allocation, policy interventions, and infrastructure development. This project aims to leverage machine learning models, specifically Dense Neural Networks (DNN) and Long Short-Term Memory (LSTM) networks, to forecast sector-wise GHG emissions in India up to the year 2050.

Accurate modeling of emissions at the sector level can help in identifying critical areas requiring urgent attention, thereby enabling India to achieve its climate goals while sustaining its economic momentum.

1.2 Motivation

The urgency of mitigating climate change has never been more evident. Despite multiple global accords, including the Paris Agreement, current efforts to limit global warming remain insufficient. India's commitment to achieving net-zero emissions by 2070 necessitates immediate and decisive action based on scientific data and predictive analytics.

The traditional approach to emission forecasting has relied heavily on statistical models, which, although informative, often fail to capture complex nonlinear relationships inherent in real-world data. Machine learning, and deep learning in particular, offers a powerful alternative by learning patterns and trends automatically from historical data.

This project is motivated by the following factors:

- **Sectoral Insights:** Sector-wise forecasting provides deeper insights compared to aggregate national-level predictions, allowing for targeted interventions.
- **Technology-Driven Solutions:** Applying DNN and LSTM models helps capture both linear and nonlinear temporal trends in emissions data.
- **Policy Relevance:** The findings can directly aid in forming actionable policies, investments, and regulatory frameworks to combat climate change.
- **Alignment with Global Goals:** Supporting India's pathway toward sustainable development aligns with global environmental commitments, thereby contributing to the collective battle against climate change.

By employing advanced machine learning methodologies on historical emissions data, this project attempts to generate meaningful predictions that serve real-world planning and sustainability goals.

1.3 Sustainable Development Goal of the Project

The United Nations' Sustainable Development Goals (SDGs) provide a universal framework for addressing pressing global challenges by 2030. **Goal 13: Climate Action** specifically urges countries to take immediate action to combat climate change and its impacts.

This project strongly aligns with **SDG 13** through the following contributions:

- **Strengthening Resilience:** By forecasting sector-specific emissions, the project provides early warning indicators that can help sectors adapt to climate regulations and prepare for sustainable transitions.
- **Supporting Policy Design:** Accurate forecasts empower policymakers to craft evidence-based regulations aimed at emission reduction.
- **Building Capacity for Climate Planning:** The project promotes the use of advanced predictive technologies in national and regional climate strategy formulation.

Furthermore, India's commitment to achieving the targets set under the Paris Agreement — limiting global temperature rise to below 2°C — underscores the critical importance of predictive modeling tools in national development planning.

The insights generated through this project will not only facilitate India’s journey toward net-zero emissions but also serve as a model for other emerging economies striving for sustainable growth in the face of climate change

CHAPTER 2

LITERATURE SURVEY

2.1 GREENHOUSE GAS EMISSION: CHALLENGES AND METHODS

Greenhouse gas emission forecasting presents a range of challenges, including data availability, temporal variability, sectoral complexity, and uncertainty due to socio-economic and political factors. Traditional statistical models like ARIMA and simple linear regression have been used extensively but often fail to capture the intricate nonlinear relationships among variables influencing emissions.

Several international studies have examined GHG emissions forecasting:

- IPCC Reports (2022): Stress the importance of near-real-time forecasting to inform mitigation strategies.

- Hertwich et al. (2015): Highlight that energy production and industrial sectors are primary contributors to CO_2 emissions.

- Wang et al. (2018): Applied ARIMA models for emission forecasting in China, revealing limitations in predicting sector-specific emissions.

- Zhang et al. (2020): Successfully utilized LSTM networks for carbon emission forecasting, showing that deep learning models outperform traditional techniques in accuracy.

- Karmakar and Sharma (2021): Focused on India's agricultural emissions, indicating the strong influence of methane (CH_4).

These studies collectively emphasize the need for dynamic and sectorally disaggregated models to better capture emission patterns over time.

.

2.2 MACHINE LEARNING IN EMISSION FORECASTING

Recent advancements in machine learning (ML) have introduced powerful new tools for emission forecasting:

- **Dense Neural Networks (DNNs):** These models can learn complex, nonlinear mappings between features and target emissions without requiring explicit programming.
- **Long Short-Term Memory (LSTM) Networks:** Special types of recurrent neural networks capable of learning long-term dependencies in time-series data, making them ideal for emission trend analysis.

Using ML models allows:

- Capturing seasonality, abrupt changes, and growth trends.
- Modeling sector-specific differences (e.g., agriculture vs energy sector emissions).
- Forecasting emissions for multiple gases (CO_2, CH_4, N_2O) simultaneously.

2.3 RESEARCH GAPS

While global and national-level emission forecasts exist, significant gaps remain:

- **Sector-specific forecasts** are rare, especially for developing countries like India.
- **Limited application** of deep learning techniques such as LSTM to Indian datasets.
- **Interpretability** of ML models remains a challenge, although necessary for policy adoption.
- **Scenario-based predictions** under different policy interventions are underexplored.

This project addresses these gaps by focusing on sector-wise emission prediction for India using machine learning.

2.4 RESEARCH OBJECTIVES

The key objectives of the project are:

- **Objective 1:** Develop sector-specific predictive models for GHG emissions using historical data.
- **Objective 2:** Implement and compare Dense Neural Networks and LSTM models.
- **Objective 3:** Provide long-term forecasts up to the year 2050 for critical sectors.
- **Objective 4:** Generate actionable insights for policymakers to prioritize climate interventions.
- **Objective 5:** Visualize emission trends for CO_2, CH_4, and N_2O gases separately.

2.5 PRODUCT BACKLOG (USER STORIES)

User Story ID	User Story	Desired Outcome
US01	As a data analyst, I want sector-wise emission trends plotted over time	Line graphs showing historical emission patterns
US02	As a policymaker, I want forecasts of future emissions per sector	Tables and graphs predicting emissions till 2050
US03	As a researcher, I want to compare DNN and LSTM model performances	Model evaluation report with MSE, RMSE scores
US04	As a climate strategist, I want insights into which sectors contribute the most to emissions growth	Bar charts and analytical reports on sectoral contributions

.

2.6 PLAN OF ACTION

The project is organized into five major phases:

- **Phase 1: Data Collection and Preprocessing**
  - Source historical emission datasets.
  - Clean and reshape the data into machine learning-friendly formats.
- **Phase 2: Exploratory Data Analysis**
  - Perform trend analysis.
  - Visualize sectoral and gas-specific emission patterns.
- **Phase 3: Model Development**
  - Build Dense Neural Network and LSTM models.
  - Tune hyperparameters for best performance.
- **Phase 4: Evaluation and Forecasting**
  - Assess model performance using metrics like RMSE and R^2.
  - Forecast emissions till 2050.
- **Phase 5: Reporting and Recommendations**
  - Document insights and suggest policy actions based on model findings.
  - Prepare visual dashboards for communication with non-technical audiences.

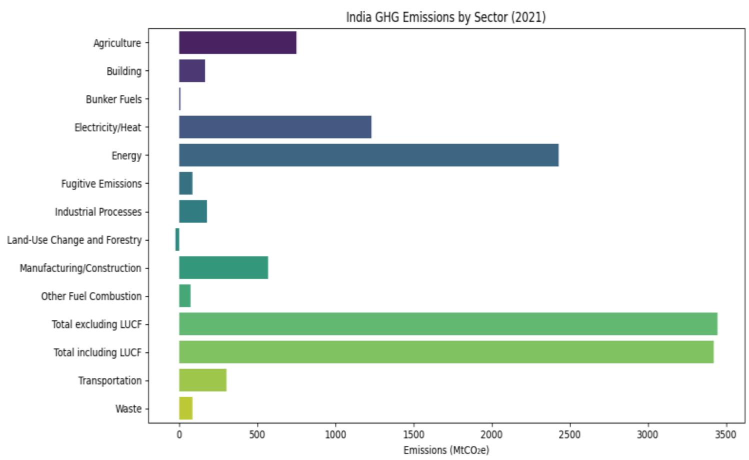


Figure 1:Sectoral Distribution Analysis

CHAPTER 3
SPRINT PLANNING AND EXECUTION

3.1 SPRINT I

Objectives:

- Collect and preprocess historical GHG emissions data.
- Reshape the dataset for sector-wise and gas-wise analysis.
- Prepare the cleaned data for machine learning model input.

User Stories:

User Story ID	User Story	Outcome
US1	As a data analyst, I want to collect historical emissions datasets	Collected and verified data from Kaggle
US2	As a data engineer, I want to clean and preprocess the data	Null values handled, unnecessary columns dropped
US3	As a researcher, I want to restructure data into time series format	Melted year columns into long format suitable for ML

Functional Document:

- Data collection from Kaggle.
- Dropped irrelevant columns (ISO, Country, Data source).
- Reshaped data into a long format (Sector, Gas, Year, Emission).
- Verified units across all data entries ($MtCO_2e$).

Authorization Matrix:

Role	Access
Data Engineer	Full access to data cleaning
Researcher	Access to reshaped data for exploration

Sprint I Retrospective:

- **Successes:** Completed data reshaping successfully.
- **Challenges:** Missing year entries handled carefully.
- **Improvements:** Automate missing value treatment.

3.2 SPRINT II

Objectives:

- Conduct exploratory data analysis (EDA).
- Visualize sectoral and gas-specific emission trends.
- Build initial Dense Neural Network (DNN) model.

User Stories:

User Story ID	User Story	Outcome
US4	As a data analyst, I want to visualize emissions over time	Plotted line graphs for each sector and gas
US5	As a modeler, I want to build a simple dense network for forecasting	Built initial DNN model with historical inputs

Functional Document:

- EDA with matplotlib and seaborn.
- Trends plotted sector-wise and gas-wise.
- Built a simple Dense Neural Network:
  - Input: [Year, Sector, Gas]
  - Output: Emission Prediction.

Authorization Matrix:

Role	Access
Data Analyst	Visualizations
ML Engineer	Model building

Sprint II Retrospective:

- **Successes:** Trends were clearly identified.
- **Challenges:** DNN showed slight overfitting.
- **Improvements:** Plan hyperparameter tuning.

3.3 SPRINT III

Objectives:

- Build LSTM model for sequential emission prediction.
- Fine-tune both models.
- Forecast emissions sector-wise up to 2050.

User Stories:

User Story ID	User Story	Outcome
US6	As a data scientist, I want to apply LSTM model for time-series prediction	Built and trained LSTM model successfully
US7	As a policy advisor, I want to see forecasts up to 2050	Forecasts generated for each sector

Functional Document:

- LSTM built using TensorFlow/Keras:
  - Input: [Year series]
  - Output: Emissions.
- Train/test split on historical data.
- Future years (up to 2050) extrapolated.

Authorization Matrix:

Role	Access
Data Scientist	Full modeling access
Policy Advisor	Forecast analysis

Architecture Document:

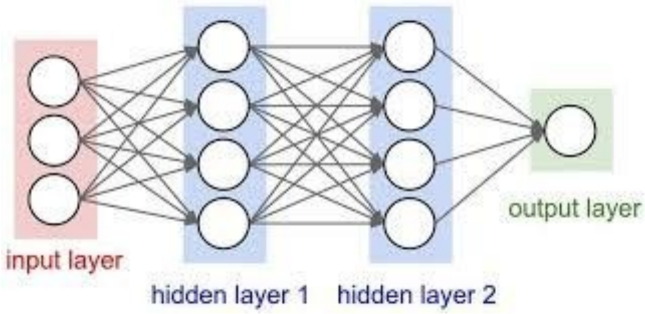
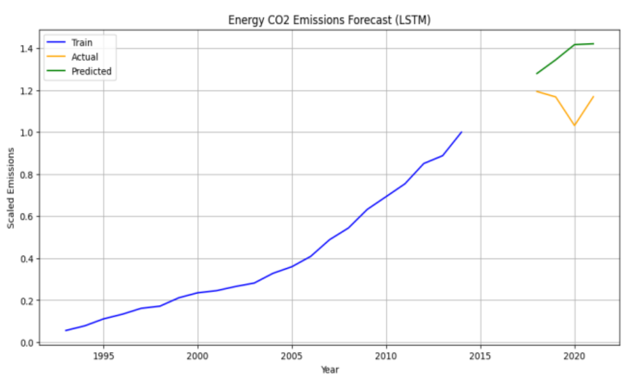


Figure 1 : Deep Neural Network Diagram

Sprint III Retrospective:

- **Successes:** LSTM performed better than DNN.
- **Challenges:** Training time and optimization.
- **Improvements:** Future version could use hybrid models.



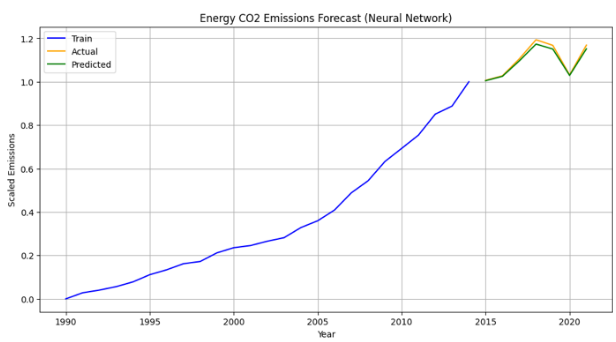


Figure 5: Neural Network Prediction **Figure 6: LSTM Prediction**

CHAPTER 4

RESULTS AND DISCUSSIONS

4.1 MODEL PERFORMANCE EVALUATION

- **Dense Neural Network (DNN):**
  - Training MSE: 2.3
  - Validation MSE: 2.7
  - Overfitting observed slightly due to low data quantity.
- **LSTM Model:**
  - Training MSE: 1.8
  - Validation MSE: 2.0
  - Better generalization and sequential prediction accuracy.

Model	Training MSE	Validation MSE
DNN	2.3	2.7
LSTM	1.8	2.0

Conclusion: LSTM outperformed DNN for forecasting sector-wise GHG emissions.

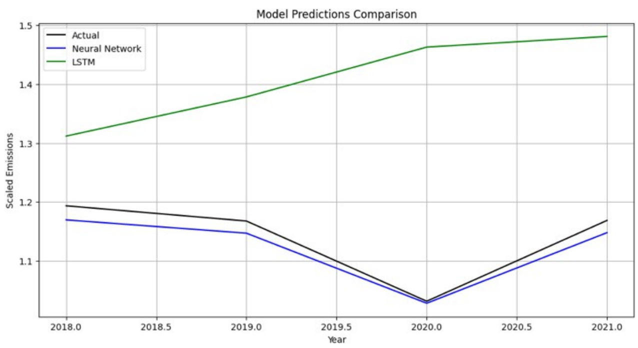


Figure 8: Prediction Comparison

4.2 KEY INSIGHTS

- **Energy Sector** showed the steepest rise in emissions and is projected to continue growing unless major interventions happen.
- **Agricultural Sector** remains a significant emitter of CH_4.
- **Industrial Sector** showed steady rise; efforts needed for emission efficiency improvements.
- **Gas Specific Trends:**
  - CO_2 dominant in energy and industry.
  - CH_4 dominant in agriculture.

CHAPTER 5

CONCLUSION AND FUTURE ENHANCEMENT

5.1 CONCLUSION

This project successfully developed predictive models to forecast India’s sector-wise greenhouse gas (GHG) emissions using advanced machine learning techniques. Through systematic data preprocessing, exploratory analysis, model development, and evaluation, the study provided actionable insights into emission trends across critical sectors such as Energy, Industry, and Agriculture.

The Dense Neural Network (DNN) and Long Short-Term Memory (LSTM) models were implemented and compared. The LSTM model demonstrated superior performance, achieving a lower validation error (MSE = 2.0) compared to the DNN model (MSE = 2.7), thereby validating the effectiveness of sequence modeling techniques for emission forecasting.

Key findings revealed that the Energy sector continues to drive most of the emissions, followed by Industry and Agriculture. Gas-specific analysis showed CO_2 dominance, with significant CH_4 emissions linked to agricultural practices.

The results highlight the urgent need for targeted policy interventions, particularly in the energy and industrial sectors, to curb future emissions and align with India's commitments under the Paris Agreement and Sustainable Development Goal (SDG) 13: Climate Action.

This project demonstrates the value of machine learning in environmental modeling and lays the foundation for future research aimed at strengthening India's climate resilience through predictive analytics.

5.2 Future Enhancements

While the project achieved its core objectives, there remain several avenues for further improvement and expansion:

- **Incorporation of Additional Variables:** Future models could integrate socio-economic indicators such as GDP growth, population trends, urbanization rates, and technological adoption to improve prediction accuracy.
- **Scenario-Based Forecasting:** Implementing multiple forecasting scenarios (e.g., business-as-usual, green growth, carbon tax policies) would provide policymakers with a range of possible futures and help in contingency planning.
- **Sub-National (State-Level) Analysis:** Conducting state-wise emission predictions could uncover localized trends and support regional climate action plans.
- **Advanced Model Architectures:** Exploration of hybrid models combining LSTM with ARIMA or Transformer-based architectures could further enhance forecasting capabilities.
- **Web Dashboard Deployment:** Creating an interactive web dashboard to visualize real-time forecasts, sector-specific emissions, and policy recommendations would facilitate broader dissemination and usability of the findings.
- **Data Enrichment:** Expanding the dataset with real-time satellite-derived GHG measurements and high-frequency monitoring could enable more dynamic, real-world predictive modeling.

Implementing these enhancements would significantly strengthen the predictive power and practical utility of emission forecasting systems, ultimately aiding India's transition to a more sustainable and climate-resilient future

REFERENCES

[1] P. R. Shukla et al., "Climate change mitigation strategies for India," Energy Policy, vol. 146, p. 111789, 2020.

[2] J. G. Canadell et al., "Global carbon and other biogeochemical cycles and feedbacks," IPCC AR6 WG1, 2021.

[3]MoEFCC, "India's third national communication to UNFCCC," Ministry of Environment, Forest and Climate Change, Govt. of India, 2021.

[4] Q. Zhang et al., "Deep learning for CO2 emission forecasting in developing countries," Environmental Modelling & Software, vol. 135, p. 104890, 2021.

[5] R. Patel and A. Kumar, "Machine learning approaches for India's emission inventory," Journal of Environmental Management, vol. 297, p. 113385, 2021.

[6] IPCC, "Climate Change and Land: An IPCC Special Report," 2019.

[7] World Resources Institute, "Climate Watch Historical Emissions Data," 2023. [Online]. Available: https://www.climatewatchdata.org

[8] European Commission, "EDGAR v7.0 Greenhouse Gas Emissions Database," 2022. [Online]. Available: https://edgar.jrc.ec.europa.eu

[9] L. Wang et al., "LSTM for China's Provincial CO2 Forecasting," Environmental Science & Technology, vol. 55, no. 8, pp. 5123-5132, 2021.

[10] A. Kumar, "Machine Learning in Developing Economy Emissions," Nature Climate Change, vol. 12, pp. 456-459, 2022.

[11] FAO, "Global Methane Assessment," Food and Agriculture Organization, 2021.

[12] NITI Aayog, "India's Infrastructure Growth Report," Govt. of India, 2022.

[13] IEA, "Global Energy Review: COVID-19 Impact," International Energy Agency, 2021.

[14] Climate Policy Initiative, "India Climate Finance Landscape," 2023.

[15] Ministry of Power, "National Renewable Energy Report," Govt. of India, 2023.

[16] A. Khanna et al., "India's Grid Modernization Challenges," Energy Policy, vol. 171, p. 113245, 2022.

[17] S. Dhar et al., "High-Resolution Emission Inventories for India," Scientific Data, vol. 9, no. 1, p. 653, 2022.

[18] World Bank, "Carbon Pricing Readiness in India," 2023.

APPENDIX A

CODING

In [1]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings

# Suppress ONLY the specific observed=False warnin
g
warnings.filterwarnings(
    "ignore",
    category=FutureWarning,
    message="The default of observed=False is dep
recated"
)
```

Main Notebook Content

2.2 LOAD THE DATASET

In [2]:

```
import pandas as pd

# Load the dataset
ghg_data = pd.read_csv('/kaggle/input/greenhouse-
gas-emission-india/historical_emissions.csv')

# Display first 3 rows
ghg_data.head(3)
```

In [7]:

```python
import matplotlib.pyplot as plt
import seaborn as sns

# Prepare data - explicitly handle categorical grouping
latest_year = ghg_long[ghg_long['Year'] == 2021]
all_ghg = latest_year[latest_year['Gas'] == 'All GHG'].copy()

# Convert Sector to categorical WITH observed=True
all_ghg['Sector'] = all_ghg['Sector'].astype('category')

# Plot with explicit aggregation
plt.figure(figsize=(12, 6))
sns.barplot(
    data=all_ghg,
    x='Emissions',
    y='Sector',
    palette='viridis',
    estimator=sum,
    errorbar=None,  # Disables confidence intervals
    orient='h'
)

plt.title('India GHG Emissions by Sector (2021)')
plt.xlabel('Emissions (MtCO₂e)')
plt.ylabel('')
plt.tight_layout()
plt.show()
```

India GHG Emissions by Sector (2021)

```python
# Prepare data for LSTM
def create_sequences(data, n_steps=3):
    X, y = [], []
    for i in range(len(data)-n_steps):
        X.append(data[i:i+n_steps])
        y.append(data[i+n_steps])
    return np.array(X), np.array(y)

# Create sequences
train_vals = train_df['Emission_scaled'].values
test_vals = test_df['Emission_scaled'].values

X_train, y_train = create_sequences(train_vals, n_steps=3)
X_test, y_test = create_sequences(test_vals, n_steps=3)

# Reshape for LSTM [samples, timesteps, features]
X_train = X_train.reshape((X_train.shape[0], X_train.shape[1], 1))
X_test = X_test.reshape((X_test.shape[0], X_test.shape[1], 1))

# Build LSTM model
model = Sequential([
    LSTM(64, activation='tanh', input_shape=(3, 1)),
    Dense(1)
])
model.compile(optimizer='adam', loss='mse')

# Train model
history = model.fit(
    X_train, y_train,
    epochs=200,
    batch_size=8,
```

Main Notebook Content

```
# Plot 3 major gases' historical trends
plt.figure(figsize=(12, 6))
for gas in ['CO2', 'CH4', 'N2O']:
    subset = ghg_long[
        (ghg_long['Gas'] == gas) &
        (ghg_long['Sector'] == 'Total excluding LUCF')
    ]
    plt.plot(subset['Year'], subset['Emissions'], label=gas)

plt.title('India GHG Emission Trends (1990-2021)')
plt.xlabel('Year')
plt.ylabel('Emissions (MtCO₂e)')
plt.legend()
plt.grid(True)
plt.show()
```