INTERNATIONAL
UNIVERSITY OF
APPLIED SCIENCES

IU International University of Applied Sciences
Master of Science in Computer Science

Modelling and Forecasting Global Renewable Energy Adoption: A Data-Driven Analysis of Causal and Predictive Patterns Using Authoritative Sources

A Master’s Thesis submitted in partial fulfilment of the requirements for the degree of Master of Science at IU International University of Applied Sciences

Name: Sunaina Manjunath
Year: 2025

Acknowledgements

I am also grateful to IU International University of Applied Sciences for providing a structured and supportive academic environment. The access to high-quality resources and mentoring has enabled me to explore an area of global importance with both depth and clarity.

Special thanks to my family and close friends for their continuous encouragement and understanding during this academic journey. Their unwavering support kept me focused and motivated throughout.

Finally, I acknowledge the open-access datasets and research repositories that made this data-driven study possible, and I hope this work contributes meaningfully to the discourse on renewable energy and sustainability.

Abstract

This thesis explores the economic, demographic and policy dynamics driving the global uptake of renewable energy technologies in the years 2000-2023. In the study, I create a harmonised multi-source country-level panel dataset and approximate a joint econometric and machine-learning approach in both explanatory and predictive directions. The methodological framework combines both fixed-effects regression to quantify structural determinants in 211 countries, time-series forecasting with ARIMA, Prophet and XGBoost on three representative case countries, and the explainability layer on SHAP values to evaluate transparency of the model.

Both the empirical findings indicate that there is a consistent relationship between economic development, population scale and electricity access in terms of increased renewable energy capacity, whereas energy intensity exhibits a negative correlation. Experiments in forecasting prove that ARIMA gives the most consistent short-term predictive value across nations, whereas Prophet is more successful in non-linear growth patterns. XGBoost is a good choice in high-variance scenarios, but more interpretability support is necessary, and SHAP analysis can offer it, which means that economic indicators continue to be the most important predictors of renewable energy growth.

The thesis finds that an econometric inference paired with machine-learning prediction can help people get a deeper picture of global renewable energy transitions. It also singles out data quality, heterogeneity of policies and generalisability of models as the major weaknesses pointing to definite directions of future studies.

***Keywords**: Renewable energy adoption, panel data analysis, forecasting, machine learning, clustering, policy analysis.*

Table of Contents

1. Introduction 1

1.1 Context and Motivation 1

1.2 Problem Definition and Research Gap 2

1.3 Aim and Objectives 4

1.4 Research Questions 5

1.5 Scope and Delimitations 6

1.6 Structure Overview 6

2. Literature Review 8

2.1 Global Renewable Energy Adoption – Trends and Statistics 8

2.2 Policy Instruments, Market Forces, and Technological Drivers in Renewable Adoption 11

2.3 Academic Models for Forecasting and Causal Inference in Renewable Energy Research 14

2.4 Machine Learning Applications in Renewable Energy Forecasting and Clustering 18

2.5 Barriers and Enablers to Renewable Energy Adoption: A Policy and Technology Perspective 21

2.6 Review of Computational Techniques in Renewable Energy Analytics 23

3. Research Methodology 26

3.1 Research Design and Philosophical Positioning 26

3.2 Data Sources and Collection Strategy 27

3.3 Data Cleaning and Preprocessing 30

3.4 Modelling Strategy and Implementation 34

Chapter 4: Empirical Results and Analysis 38

4.1 Overview of the Analytical Pipeline 38

4.2 Descriptive and Exploratory Results 39

4.3 Econometric Results: Panel Regression 41

4.4 Forecasting Results 47

4.5 Explainability Results 50

Chapter 5: Discussion and Interpretation 53

5.1 Linking Econometric Results to the Literature 54

5.2 Causality versus Prediction: Methodological Reflection 55

5.3 Interpretation of Forecasting Outcomes 56
5.4 Cluster Interpretation through Innovation Diffusion Theory 57
5.5 Explainability and Policy Relevance 58
Chapter 6: Conclusion, Limitations, and Future Work 60
6.1 Summary of Key Findings 60
6.2 Contributions to Knowledge and Practice 62
6.3 Policy Implications 63
6.4 Limitations 63
6.5 Future Research Recommendations 64
6.6 Concluding Remarks 65
References 67
Appendix 73
A1. STEP 1: Environment setup & data loading 73
A2. Step 2: Dataset inspection & variable selection 73
A3. STEP 3: Missingness Diagnostics & Country Filtering 75
A4. STEP 4: Integration with World Bank (WDI) Indicators 76
A5. STEP 5: Missing-Data Treatment & Imputation Rules 79
A6. STEP 6: Exploratory Data Analysis (EDA) 81
A7. Step 7: Econometric Modelling (Panel Regression) 85
A8. Step 8: Forecasting 89
A9. Step 9: Explainability (SHAP analysis for XGBoost) 94
Declaration of Authenticity 103

Table of Figures

Figure 1: Global electricity installed capacity from 2000 to 2023 1

Figure 2: Barriers To Renewable Energy Adoption 3

Figure 3: The performance of the world's countries in achieving the goal of transitioning to renewable energy 8

Figure 4: Gross electricity production from renewables 9

Figure 5: Worldwide investment in renewable energy 10

Figure 6: Process of innovations 13

Figure 7: Flowchart of Arima and Machine Learning for time series forecasting 15

Figure 8: ML types and algorithms 18

Figure 10: Forecast Comparison - Germany 49

Figure 11: Forecast Comparison – India 49

Figure 12: Forecast Comparison – USA 50

Figure 13: SHAP summary plot 51

Figure 14: SHAP dependence plot 51

List of Tables

Table 1: Dataset Summary Table 29

Table 2: Final Dataset Schema Table (Variables, Units, Transformations) 34

Table 3: Descriptive Statistics Output 40

Table 4: Top Countries' Output 41

Table 5: Pooled OLS summary Tables 43

Table 6: Fixed Effects summary 44

Table 8: FE vs RE comparison table 46

Table 9: VIF Output 46

Table 10: Forecast accuracy table 48

Table 11: Research Objective / Question Mapping Table 62

Abbreviations

ARIMA – AutoRegressive Integrated Moving Average

GDP – Gross Domestic Product

IRENA – International Renewable Energy Agency

ML – Machine Learning

PCA – Principal Component Analysis

1. Introduction

1.1 Context and Motivation

The worldwide pressure to go renewable with alternative sources of energy has continued to swell over the last ten years due to fast-paced changes in climate change, exhaustion of natural resources, and the growing energy demands. Since the 1.5 °C climate target in the Paris agreement continues to penetrate the strategies of governments, institutions, and industries, renewable energy, and notably, solar and wind energy, has come to be the basis of decarbonization policies (Brecha et al. 2022). Such sources not only provide sustainability in the long term but also lead to an increase in energy security, lessening geopolitical risks, and provide green jobs.

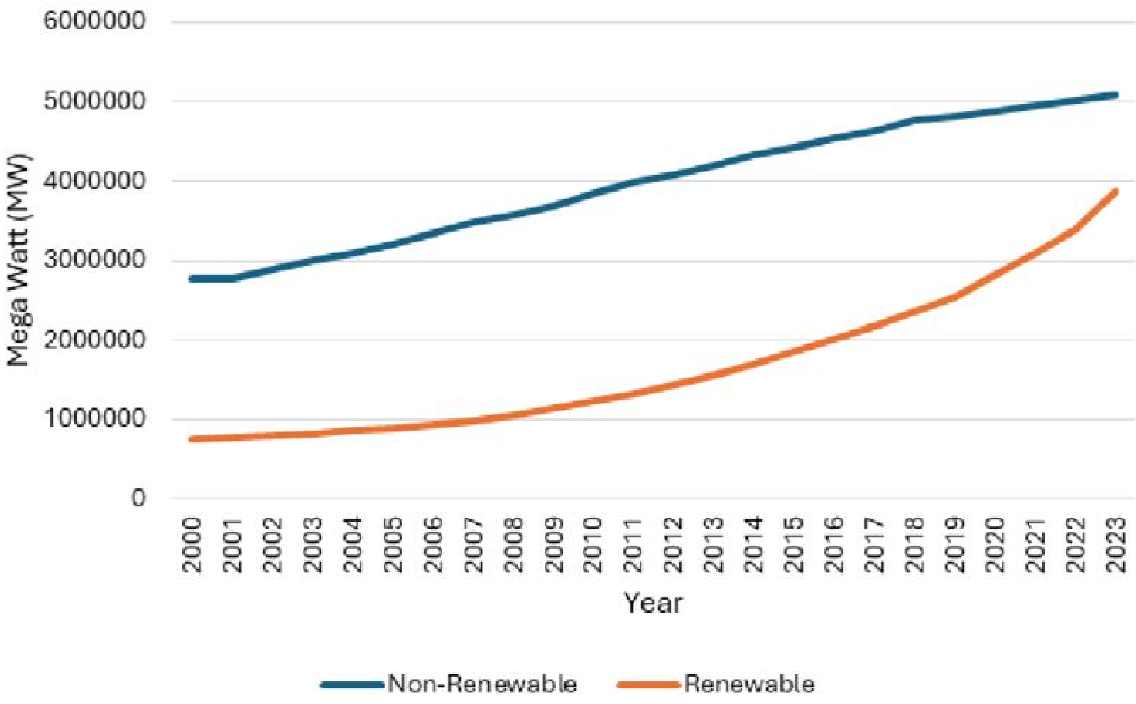


Figure 1: Global electricity installed capacity from 2000 to 2023

(Source: Saeed & Siraj, 2024).

The figure statistically shows the installation of electricity capacity trend across the world from 2000 up to 2023 in comparison to both the renewable and non-renewable sources. Although the dominance of non-renewables persists, the upward trend of renewables is steep, indicating the rapid increase in their adoption. Such a visual emphasizes one of the factors urging the transition to energy, its movement, and supports the definition of the study in context.

Renewable energy has experienced great improvement in developed and developing economies alike. From 2010 to 2022, the capacity of solar photovoltaic (PV) increased more than ten times globally, whereas the capacity of wind power increased to over three times (Swadi et al. 2024). The trends of adoption, though this is a momentum, do not follow the same pattern. Although the restricting policies, mechanisms of financing, technological accessibility, and geography influence ED in different ways, the countries with comparatively high absolute installed capacities are China, the United States, and Germany (Crijns-Graus et al. 2022). These trends need to be explored, using data and in their granular form, to reveal the subtle differences within adoption uncovering.

Such exploration is now provided on a rich basis by open-access data repositories. Such initiatives as the International Renewable Energy Agency (IRENA), the World Bank, the Energy Institute, and REN21 present annual country-specific data on installed capacities, that is, investment flows, as well as on energy policy indicators. These datasets can be used together with analytical software, such as Python and machine learning algorithms, to create powerful, scalable, national- and region-level energy transition evaluations. Simultaneously, visualization tools and open-source dashboards have made energy analytics more accessible and revealed the trends to governments, researchers, and the population to interact in real-time (Osamika et al. 2025)

Nevertheless, previous studies have investigated macro-level adoption, portfolio trends, or policy assessment, but fewer studies have taken on a united, computational method to integrate statistical inference, trend forecasting, and unsupervised clustering between nations. Additionally, econometric analysis and simulation models are not the only methodological groups of literature variables; that is why it is frequently not possible to compare or repeat them. The proposed thesis is an attempt to fill this methodological hole with a reproducible and open-source computational workflow to understand global adoption trends based on authoritative data and powerful statistical techniques. This study will incorporate causal inference and predictive modelling, unlike pure descriptive models, to define the fundamental drivers and future trends in renewable adoption.

This incentive is based not merely on the academic importance, but also on the functional importance. To a policymaker, a better comprehension of what works where and why can be used in the formulation of specific incentives and enforcement mechanisms. In the case of researchers, reproducible methodology founded on publicly available datasets and codes is added to the emerging science of energy informatics. And with respect to global stakeholders, such as intergovernmental organizations, as well as local utilities, benchmarking patterns of adoption and making predictions is value-added to cross-country cooperation and forward-looking infrastructure.

This study is, in fact, driven by the convergence of global need, availability of data, computational possibility, and agency. It intends to offer a policy-relevant, insightful, and replicable contribution to the current world debate on renewable energy transitions by relying on the credibility of its sources and computational models.

1.2 Problem Definition and Research Gap

Although there have been significant advances in the use of renewable energy worldwide, the adoption of clean energy is still uneven, broken, and insufficiently comprehensible regarding its causal processes. Although countries such as Denmark and Germany have attained a high level of penetration of solar and wind energy as compared to the overall generation, there are many low and medium-income countries that are lagging despite their favorable geographical locations and declining technology prices. This imbalance underscores the serious failure in our knowledge of the structural, policy, and economic factors that either serve to enhance or restrain the adoption of renewable energy in different regions.

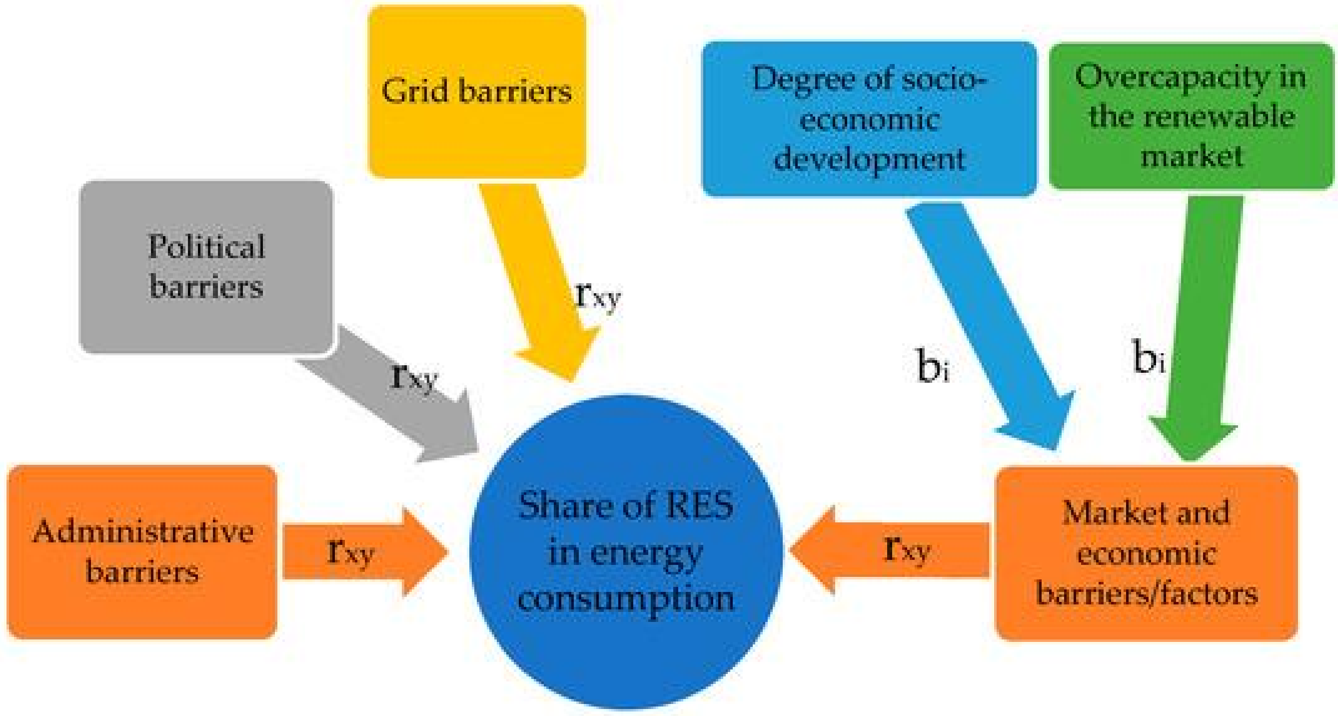


Figure 2: Barriers To Renewable Energy Adoption

(Source: Gajdzik et al. 2023)

A lot of available studies have been limited to single countries' case studies or at higher ranks, comparative statistics that do not go further in providing further empirical evidence on how growth and evolution factor differently with time, region, or are simply unified by a clear driver of growth or development. Most reports of international organizations, including IRENA and REN21, give descriptive analytics in detail, but most are not dynamic, not prospective, and not presented as formal models (Koutoudjian et al. 2021). Furthermore, various national reports on energy transition are unfocused in terms of methodological transparency, or they use inaccessible data pipelines, making it difficult to use such reports in comparative, scalable, or predictive analysis.

In the academic field, several literature reports have determined the effect of policy instruments, e.g., a feed-in tariff, renewable portfolio standards, and green subsidies, on adoption. Migration flows, as well, however, are sometimes examined in individual components but fail to take into consideration the multivariate interaction or cross-country heterogeneity (Soto Nishimura & Czaika, 2024). Little use is also made of machine learning or data-driven modeling methods to recognize latent trends, growth predictions, or to cluster countries based on their structural form. Besides, it usually lacks reproducibility, which is a critical criterion in computational research, with data unavailable online, pipelines uncharacterized, and inconsistent application of statistical measures of evaluation.

There are three major gaps identified in this thesis. One, the lack of a unified, reproducible formula of analysis that consists of both statistical and machine learning to examine the adoption of renewable energy across the world. Second, the level of authoritative primary data incorporation (e.g., IRENA, World Bank, Energy Institute, etc.) into the scalable computational models is lacking. Third, the current literature does not have policy-conscious clustering with the ability to sort countries not only according to absolute adoption, but also according to similarities about GDP per capita, policy strength, or geographic potential.

The consequences of these gaps can be experienced in the real world. These officials are left with disjointed evidence, and thus, benchmarking the progress and interventions is unlikely to become feasible. It presents methodological inconsistencies between academic researchers to allow them to generate cumulative knowledge. And the global watch bodies do not have a comparative, empirical picture upon which to keep track of the world as progressing in accordance with the commitment to climate.

The challenges in this thesis will be discussed by developing a data-based, reproducible, and policy-considerate framework to examine the rise of solar and wind energy adoption between the years 2000 and 2023 across countries. It is going to integrate statistical inference (e.g., panel regression), predictive analytics (e.g., ARIMA, Prophet), and unsupervised learning (e.g., K-means clustering, PCA) to produce actionable data. By so doing, it will serve not only as a contribution to the scholarly discourse of energy transitions but also in the field of actual policymaking and sustainability monitoring in the global scene.

1.3 Aim and Objectives

The main purpose of this study will be to construct a conceptual, data-based, replicable, analytical model that will examine, modelize, and forecast the worldwide trend of renewable and clean energy adoption using solar and wind energy. The framework will be used to analyze the time, space, and policy-induced factors of influence in adoption across countries between the period 2000 and 2023 using verified datasets and advanced algorithms. By so doing, the thesis is intended to transform itself into a beyond descriptive statistics step, to add insights that would be meaningful and relevant to policy discussion, to the international slide to renewable energy.

1. To form a homogenous and replicable dataset.

The aim is to aggregate, clean datasets of multiple sources, which are the primary ones, that include IRENA, the World Bank, REN21, and the Energy Institute, to form a longitudinal and country-based panel covering the adoption of solar and wind energy, policy indicators, socio-economic factors, and the social and economic mix of energy. The utilization of open-access Kaggle repositories can be applied only in their additional verification or visualization, but not as a primary source.

2. To compare the time and place adoption series.

It will comprise descriptive statistical summaries, visualization, and trend lines to research the growth trend of solar and wind capacities within countries and regions. Python visualization tools (e.g., Plotly, Matplotlib) to create interactive time series will be used in forming a basic macro-level definition of the dynamics of the adoption.

3. To carry out the basis and predictive modeling.

The overall goal involves the use of quantitative methods, which include:

Fixed effects or random effects panel regression models to determine the causal effect of the variables on the growth of renewable capacity (GDP per capita, population, policy support, and CO_2 emissions)

Predictive analytics like ARIMA, Exponential Smoothing, or Prophet to forecast adoption rates in selected countries and confirm the validity of prediction scores with time-based metrics of cross-validation (e.g., RMSE, MAE).

4. To perform clustering and dimensionality reduction.

The countries will be clustered into a common set of adoption behaviors, the intensity or structural attributes, using the unsupervised machine learning strategies, including the K-means clustering, Hierarchical Clustering, and Principal Component Analysis (PCA). This will give a typology of adoption archetypes, a case of leaders, laggards, and new enterprise adopters.

5. To draw inferences about the policy through comparative analysis.

Using the models and clusters, the thesis shall derive comparative policy implications, finding the most significant drivers and the most common barriers. The results will be connected to the global policies like SDG 7 (Affordable and Clean Energy), the Paris Agreement, and concrete evidence-based proposals will be provided.

The goals are consistent with the academic requirements of a 120 ECTS thesis in IU Computer Science. Data engineering, machine learning, statistical modelling, and policy analysis are included, all in a reproducible workflow in Python. The computational rigor and the societal value of renewable energy analytics are set to be added by integrating technical aura with real-life policy applicability to the fields of research effort.

1.4 Research Questions

To be sure of methodological coherence and academic rigor, the formulation of specific, testable, and computationally feasible research questions is necessary. The thesis is supported by 5 research questions, which are interrelated and are directed at addressing each of the elements of the global renewable energy transition.

- **RQ1**: How are the global practices in touring and embracing solar and wind energy evolving over the period between 2000 and 2023?

The purpose of this question is to have a historical score as well, to determine the trends of renewable energy capacity growth and its spatial dispersion. It would entail descriptive statistics and time-series visualization, and would provide a background basis to more intensive modeling.

- **RQ2:** How are the policy, economic, and demographic factors the major drivers of the adoption of renewable energy in countries?

In this case, the objective is to implement panel regression equations to be able to test the causal effect of predictors in the form of:

- Policy scores (renewable energy policy, e.g., FITs, tax credits)
- Urbanization rates and GDP per capita.
- CO_2 emissions and fossil addiction.

This will aid in testing that there is a statistically significant effect of certain factors over time and among entities.

- **RQ3**: Are time-series models prone to prediction adoption patterns, and in what ways can these models be applied throughout the varied contexts of countries?

In this question, it is necessary to evaluate predictive performance based on such models as ARIMA, Prophet, and exponential smoothing. The determination of generalizability and temporal reliability will be conducted with the help of model evaluation metrics (e.g., RMSE, MAE, MAPE).

- **RQ4**: In which ways can countries be grouped into typologies of their adoption behaviors, policy environment, and structural variables?

This shall be handled with the help of unsupervised machine learning. The clustering analysis will reveal latent similarities, which will bring a new perspective on interpreting adoption typologies (e.g., early adopters, slow movers, overperformers).

- **RQ5**: What were (are) the policy hints and comparative learning that are made with the most and the least successful nations in pushing the renewable energy development?

The last research question is an integration of all the previous findings towards a policy-relevant summary. It will draw comparisons between countries that are doing well and those that are performing poorly to derive the best practices, ubiquitous obstacles, and country-specific advice.

The joint combination of these research questions results in a multi-layered form of research question: macro-pattern recognition (RQ1), causal identification (RQ2), predictive modeling (RQ3), structural clustering (RQ4), and lastly, in applicable implications (RQ5). These are correlated with the thesis goals and plunge into the academic texts as well as into the practical requirements of the policies.

1.5 Scope and Delimitations

The thesis seeks to establish data-driven models to examine the growth in the utilization of solar photovoltaic (PV) and onshore/offshore wind energy between 2000 and 2023 on a global scale. It is devoted to discussing historical and predictive dynamics, determining the main adoption motivators, and distinguishing typological varieties of adoption in the countries. The analysis is solely based on quantitative analysis, such as descriptive statistics, panel data analysis, time-series models, and machine learning-based clustering, which may be performed with Python because of its transparency, scalability, and reproducibility.

The data analysis is based on country-level data obtained via IRENA, the World Bank, REN21, and the Energy Institute on such indicators as installed renewable capacity (MW), GDP per capita, population, policy instruments, and carbon emissions. Hydroelectricity, geothermal, and biomass are not included as there is no consistency in data, and the rate of innovation is slow, which ensures consistency in comparing observations.

Only publicly accessible secondary data is used in the study, with data imputation and filtering used to fill gaps, but there are still a few limitations. It does not involve qualitative studies, project-based, and firm-based studies, but rather a macro-level approach to build policy relevance. Short-term (2025-2030) trend projections are limited by the availability of data. The research methodology is rigorous and feasible, focusing on scalable and replicable findings on the adoption of renewable energy in the world.

1.6 Structure Overview

The structure of this thesis is designed based on the IU Thesis Handbook (2024), which guarantees the coherence, academic depth, and transparency of the methodology. The complete thesis will be about 24000 words, structured in six main chapters, with the emphasis on quantitative and policy-significant and reproducible computational analysis.

- **Chapter 1: Introduction** - The chapter is the context setter of the study since it justifies the context of the study and provides motivation, research gap, objectives, and details of the study, including research questions. It also gives the scope and limitations, which specify the geographical, methodological, and technological limits of the research. The structure summary alone is a reflection of the thesis roadmap, which is a conclusion foreseeing the end of the chapter.

- **Chapter 2: Literature Review** - The second chapter offers a critical review of the literature available concerning renewable energy acceptance in scholarly literature and policies. It is subdivided into sub-sections, which include:
  1. Renewable diffusion and innovation adoption theoretical underpinnings,
  2. Frameworks of Policies (e.g., feed-in tariffs, green subsidies),
  3. These are cross-country studies that have shown empirical results in the past,
  4. Industrial ways of doing computations in energy analytics (e.g., ML, forecasting),
  5. Identified research gaps.
  6. The literature review provides an excellent academic basis and ensures that the thesis has a place in the existing arguments.
- **Chapter 3: Methodology** - The chapter confirms the details of the research design, data sources, methods of data cleaning, modelling measures, the environment used to run the Programme (Python), and protocols associated with reproducibility. Each of the methods, panel regression, time-series forecasting, and clustering, will be elaborated in subsections and their role in answering questions posed by the research. It also deals with the evaluation metrics (e.g., RMSE, R 2), validation approaches, and any ethical concern (e.g., data license, and privacy), e.g., GDPR conformance or alignment.
- **Chapter 4: Findings and Analysis** - In this case, the results of both approaches to analysis are reported and discussed. Visualization of time-series graphs, regression outputs, cluster diagrams, and forecast models is explained. All outcomes are traced to continually reposing research questions. An appropriate figure and table will be provided in this chapter and will conform to reproducible codification.
- **Chapter 5: Discussion** - The discussion is made up of a synthesis of all the analytical outcomes and the contextualization of the results in the global energy policy, sustainable development, and literature. The major themes are: the factors (country-level) that have been identified as different in their adoption; how the predictive models cope with the existing trajectories; and what typologies arise in the clustering. Critical evaluation of limitations, robustness checks, and the method caveats will also be critically evaluated.
- **Chapter 6: Future Research and Conclusion** - The concluding chapter examines the objectives of the thesis again, summarizes the main findings, and states the policy implications of the work to governments and international organizations. It also provides suggestions to future studies, including the addition of qualitative policy analysis, more subnational data, or real-time satellite data analytics.

2. Literature Review

2.1 Global Renewable Energy Adoption – Trends and Statistics

The past twenty years have seen a rampant increase in the uptake of renewable sources of energy, particularly solar photovoltaic (PV) and wind. The ever-growing concerns in climate change, air pollution, energy security, as well as the rise in competitiveness in costs of technologies, have led to policy as well as investment towards clean energy systems. This part concerns itself with the quantitative patterns in renewable energy installation, the amount of power added, electricity production proportions, and by region, based on the available literature such as the International Energy Agency (IEA), IRENA, REN21, Energy Institute / Ember, and others, containing a speculation of peer-reviewed literature.

IRENA forecasts that the additions to global power capacities based on renewable sources will reach 582 GW in 2024, a 19.8% increment over 2023. This growth was largely dominated by solar PV, which added approximately 452 GW (78%) of the additions, and wind, which made additions of approximately 114 GW. This spurt increased the cumulative worldwide installed renewable power to approximately 4,443 GW by the conclusion of 2024.

Asia has been experiencing great growth rates. Asia Pacific, headed by China, India, and southern Asian countries, had some of the fastest capacity increases. As an example, over 60% of the global PV capacity expansion of 2024 was in China alone, and a comparably high portion of wind installations.

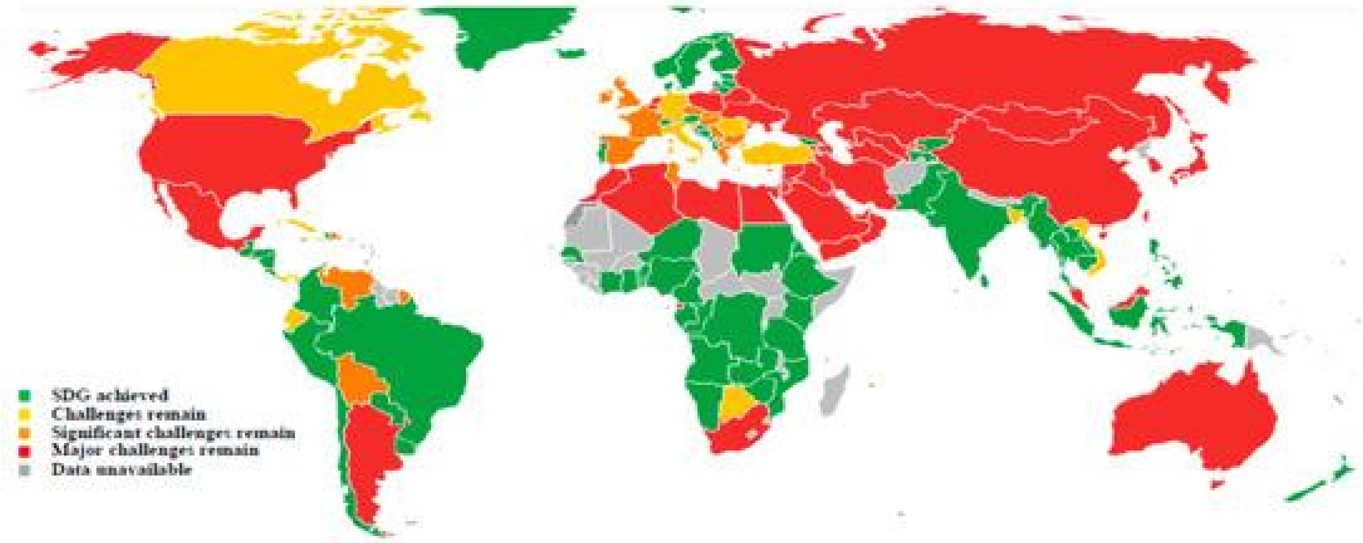


Figure 3: The performance of the world's countries in achieving the goal of transitioning to renewable energy

(Source: Mohamed et al. 2024)

According to Mohamed et al (2024), this world map shows how various nations are moving in the process of increasing the capacity of renewable energy transition.

In 2024, clean electricity generation (renewables + nuclear) accounted for over 40% of global power output for the first time. Clean power is spreading mainly through the deployment of clean energy and the fast growth in solar PV. In three years, the amount of solar generation has more than doubled, making it the fastest-growing source for electricity globally. It should be noted, however, that wind generation also rose significantly, if from a somewhat lower base. In 2024, when it comes to additional generation capacity as well as new generation, PV wind together exceeded hydropower according to Ember / Global Electricity Review.

By 2030, an outlook released on March 24th believed that renewable electricity will represent 46% of total global generation. In this projection, solar and wind for now make up 30%, though considerably less in 2023.

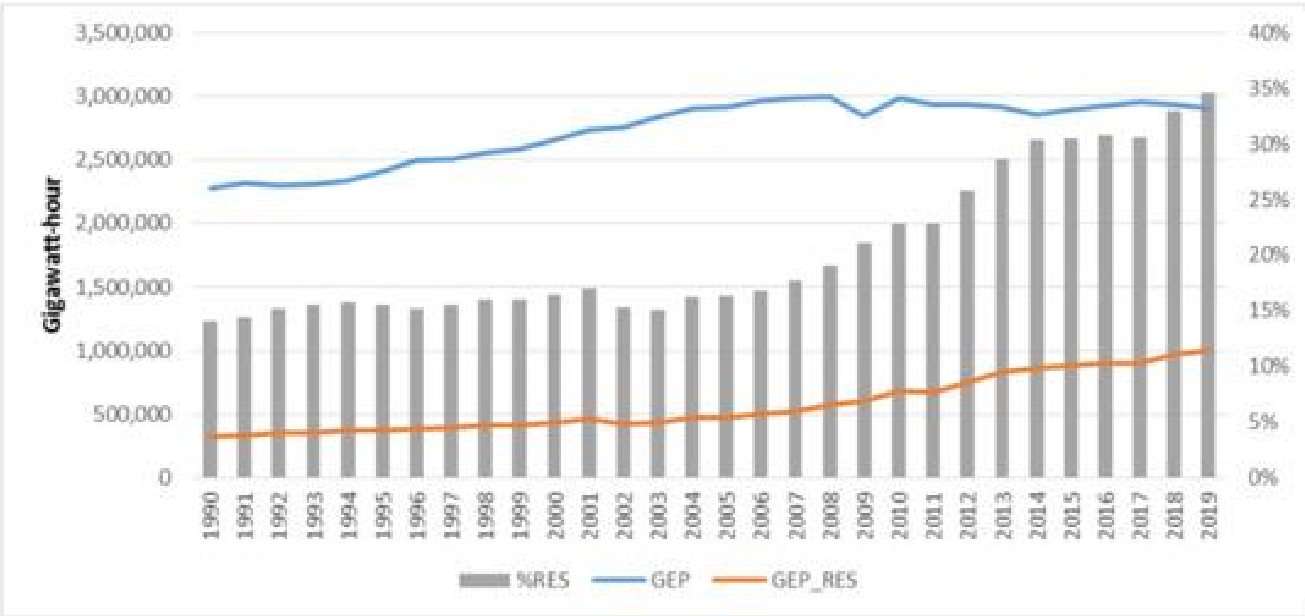


Figure 4: Gross electricity production from renewables

(Source: Matuszewska-Janica et al. 2024)

According to Matuszewska-Janica et al. (2024), the chart shows the increase in electricity production from 1990 to 2019, alongside a rise in renewable energy contributions. The share and production from renewables demonstrate consistent upward growth over the years.

The last few years have witnessed solar PV soaring ahead of all other technologies. The strong deployment has been enabled by the dramatic cost cuts in the production of PV modules, the advancement in the inverter construction technology, economies of scale, and favourable policies by various governments. Solar PV has been added to a significant extent in Asia, and especially in China.

Solar has a gradually increasing rate of growth, whereas wind energy (onshore and offshore) is expanding at a slower pace. Considering the geographical, regulatory, and grid integration opportunities, wind suffers more hitches in most regions, slowing down acceleration compared to solar. Nonetheless, there is an improvement in turbine design, offshore wind capacity, and storage integration, which are aiding.

Global trends are strongly upward, and regional disparities are large:

i) Asia (particularly China and India) leads new installations in both wind and solar. China in 2024 accounted for the bulk of the PV and a significant portion of the wind capacity increases.

ii) The European region is still the leader in cumulative installed capacity, sophisticated policy regimes, and deployment of offshore wind, but trails Asia in market share of new capacity addition.

iii) North America is growing steadily on the back of growing wind and solar projects, but is plagued by policy uncertainty in some states (e.g., permitting, grid connection) as barriers.

iv) Africa, Latin America, and the Middle East have mixed trajectories: Latin America is picking up steam, particularly in terms of deployment of solar energy; a few areas in Africa are very prospective but lag in deployment due to financing, infrastructural, and regulatory challenges. REN21's 2025 Global Status Report highlights these regional variations and emphasises that while capacity is growing everywhere, the pace and scale are very uneven.

The rate at which renewable power output ratio units is being adopted is picking up due to a combination of favourable policies, declining costs of solar and wind, and a boost in flow investment with special reference to the perspective of solar power output ratio units. Regular actions by the government, such as preferring tax cuts and government procurement requirements, boost investor confidence. Simultaneously, renewables are becoming more and more reliable due to technological advances, including efficient turbines, improved solar modules, better storage, and prediction methods, with the perspective of permitting all countries to shift sooner towards cleaner and more long-lasting sources of power output ratio units.

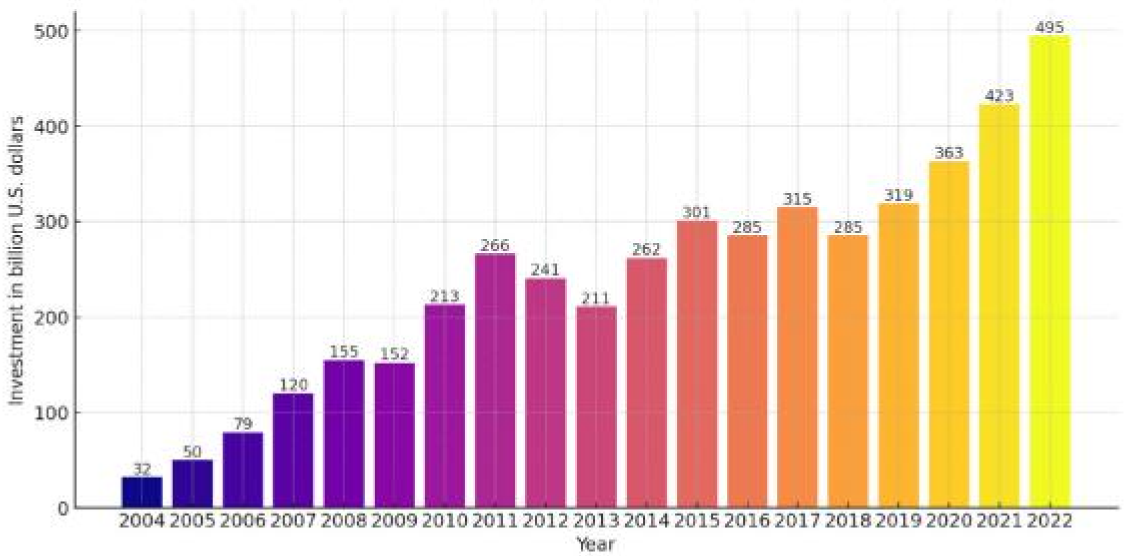


Figure 5: Worldwide investment in renewable energy

(Source: Shah, 2024)

According to Shah (2024), the graph shows the rise in worldwide investment in renewable energy from 2004 to 2022. It reached nearly $500 billion in 2022. This trend reflects the growing global commitment to the clean energy transition.

The adoption of renewable energy is challenged despite its excellent growth. Ridett & Anderson (2024) clarify that intermittent supply can hardly be balanced due to grid deficits and grid capacities. Altering subsidies, halting permits, and alternative policies of rendering things unstable are some of the things that are likely to scare away business investors. Uncertainties are linked with missing data, especially in recent projects or off-grid projects. The upfront grandiose expense, coupled with minimal financing, is a blowback, more so within the developing localities. Finally, the fact that poor interconnection, storage, and demand control are a slowing factor to integrate variable renewables is the last to consider.

Renewable energy is still resurging around the globe, and 2024 is projected to record the highest rate of power generation capacity. Renewable power is provided together with nuclear power in clean energy, supplying more than 40% of global power, which is made up of solar and wind energy. A good part of this growth is attributed to Asia, with other areas growing at dissimilar rates. Heavy policies, declining prices, and technology are knocking down the doors, yet issues such as bad infrastructure, lack of funding, regulatory obstacles, and data shortage continue to leave parts of the world behind. A combination of these dynamics exhibits the opportunities and complexities of the worldwide transition to clean, cheap energy and prepares for any additional examination by modelling, predicting, and policy review.

2.2 Policy Instruments, Market Forces, and Technological Drivers in Renewable Adoption

A complicated combination of different policy tools, market dynamics, and technological possibilities will support the large-scale implementation of renewable energy all over the world. Cost online variability and innovation lead to supply-side dynamics, but the policy frameworks form facilitating conditions of the process of diffusion. It is a synthesis of the particular academic and institutional literature on the influence of these mutually reinforcing forces on the adoption patterns based on comparative cross-country research and empirical modelling specifics.

Here, the policy tools are feed-in tariffs, renewable portfolio standards, tax incentives, and competitive auctions, which are interplayed with market conditions to produce their effect on investment choices and risk perceptions. Learning curve and economies of scale further improve costs with time and imply a positive feedback loop that increases the speed of adoption. These drivers, however, are effective in different countries depending on the capacity of institutions, the growth of financial markets in countries, and the maturity of energy systems. Empirical analysis that cross-country offers thus offers a win-win perspective on finding systematic patterns in keeping with the heterogeneity in national situations. These dynamics can only be understood to devise policy frameworks that are context sensitive and scalable, which will facilitate a long-term increase in the deployment of renewable energy.

The powerful body of research investigated the effect of policy interventions on the creation of renewable power output ratio units. They may be roughly measured as price-based, quantity-based, and market facilitation instruments.

The input salt density in Tariffs (FiTs) has traditionally been one of the most powerful price-related mechanisms. They ensure there are set payments per unit of renewable electricity fed into the grid, and therefore, revenue is guaranteed from the perspective of the investors. According to the perspective of observational research done by Coronas et al. (2022) and Garrido-Herrero et al. (2024), there are powerful positive relationships between FiTs and Solar PV/wind deployment that mainly occurred in the process of the 2000s in Europe. Yet, the issue of cost overburden and overcapacity prompted most countries with the perspective of turn with the perspective of with the help of auction-based systems.

Renewable Portfolio Standards (RPS), or Renewable Power Output Ratio Units Obligations (REO), are targets that are quantitative in nature and demand that electricity suppliers obtain a specified quota of their power output ratio units supply through renewables. They are more eminent in the United States and in some regions of Asia. Compared with the perspective of FiTs, RPS can have price competition, although it can lead to market volatility (Vo, 2024).

Opposite direction auctions, which are a segment of the country after finding their way into the world, can be used with the perspective of finding the price by making selections according to the perspective of the lowest bidders seeking renewable supply contracts. Kalvenes & Toftegaard's (2023) research indicates that auctions have resulted in irregular price reductions in the solar and wind industry in such countries as India, Brazil, and South Africa.

The instruments on top are complemented by tax incentives, subsidies, and capital grants in early-stage markets. However, as emphasized by Azizah & Amalia (2025), all of them are frequently used on behalf of the ability to accomplish goals based on institutional maturity, openness, and clarity, and investor user confidence.

The project has been led with the perspective of declining Levelized Cost of Electricity (LCOE) in many geographical areas. It shows renewables comparable with the perspective of non-subsidised financial prudence. As IRENA (2023) reports, the global weighted-average LCOE of solar PV on large scales and onshore wind decreased by 88% and 69% in 2010-2022, respectively. These attacks were found by supply chain economies, and balance of system optimisations.

There are also trends of investment in the sector by the private sector, which is affected by the willingness to invest. According to the perspective of the Khan et al (2023) report, worldwide investment in solar power output ratio units went USD 495 billion in 2022, with high payments in solar PV. Such inflows of capital are, however, tilted towards the OECD and the virtues of middle-income nations. Weaker countries tend to have greater threat premiums, reduced access from the perspective of credit, and immature capital markets, which severely restrict their ability to roll out renewables.

Notions of grid access, curtailment risks, and power purchase agreements (PPAs) impact the bankability of projects. Renewable generation can be curtailed or stranded in countries where the infrastructure system of the grids is weak or the transmission losses are high (Zhanget al. 2021).

The national renewable power output ratio units, targets, and procurement plans are also influenced by the demand side elements, preferring the patterns of electric consumption and urbanisation, as well as the rates of electrification. Although a country is considered a high-income country, and the growth of electricity demand is much slower, whereas in emerging economies, there is an equivalent demand coupled with the possibility of leapfrogging programs dramatically into cleaner systems.

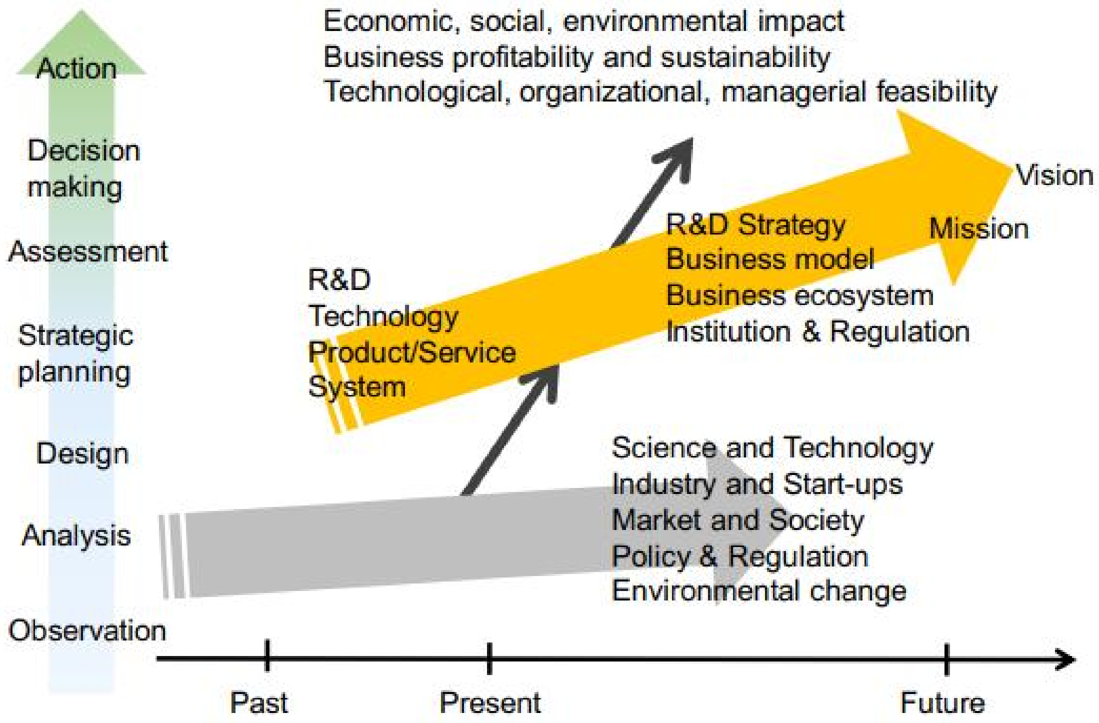


Figure 6: Process of innovations

(Source: Makowski & Kajikawa, 2021)

According to Makowski & Kajikawa (2021), this diagram shows the process of innovation, moving from past observations to future-oriented missions. It highlights the technology, business models, regulations, and societal factors in sustainable innovations.

The fastest innovation has been in solar PV. The development of first-generation crystalline silicon technology into thin-film and then perovskite technologies depicts an efficiency and longevity increase of two factors every twenty years. The National Renewable Energy Laboratory (NREL) maps this trajectory extensively, and the study by Brinkman et al. (2021) reviews this trajectory. Output has been enhanced with Inverters and Tracking systems.

Wind technology has advanced in terms of the height of turbines, the diameter of the rotor, and digital optimisation. The European cable, Russia, China, and the UK are different wind sources. Increased shares of renewables in power systems are being enabled by digitalisation and the technologies of grid integration, which comprise IoT-enabled sensors, state-of-the-art forecasting models, and battery storage. The above studies were conducted by peer-reviewed sources (Kacena et al. 2024). It focuses on the role of smart grid infrastructures, demand response, and virtual power plants in facilitating renewable variability (Abdelkader et al. 2024).

The distributed generation is on the increase because of consumer choice, resilience requirements, and cost.

Although the aspects presented above have a general impact, their impacts do not have a similar impact across nations. Many researchers have pointed to the significance of institutional quality, governance, policy stability, and grid maturity to identify the success of the renewable energy start-up (Zia et al. 2022). One could use the case of FiTs in Germany, where transparency in the laws worked and bank entails worked well, but not in Spain, where policies are made retroactively.

It is also due to energy dependency. More life-dependent countries, such as those with heavy reliance on fossil fuel exports, are at risk. It can face political opposition or the lack of economic incentives to make the change. In contrast, a renewables level will progress more in energy-importing countries (e.g., Japan, South Korea) within their energy security strategies.

There is growing academic consensus that multi-level and cross-sectoral governance, national requirements with regional/local practice, and stakeholder involvement are needed (Gargano, 2021).

There is a mix of policy instruments (FiTs, RPS, auctions), market (costs drop, investment), and technological innovation (PV/wind improvements, digital integration) spread behind the issue of the adoption of renewable energy.

- The key elements in facilitating widespread diffusion are policy design, institutional quality, and financial de-risking.
- Already, the world literature demonstrates high levels of heterogeneity in its results depending on the level of governance quality, the rate of income, and the preparedness of infrastructures.
- These variables need to be incorporated in future research in quantitative studies in a bid to describe the direction of adoption and predict future shifts.

This overview has set the analytical and methodological decision-making of this thesis. In the following part of this paper, it will visualize scholarly input on the forecasting, clustering, and causal inference of renewable adoption to provide the theoretical foundation of the modelling framework in Chapter 3.

2.3 Academic Models for Forecasting and Causal Inference in Renewable Energy Research

The growing complexity of reconfiguring renewable power output ratio unit systems all over the world has led to the perspective of the creation process and utilisation of advanced forecasting models. These academic facilities have two logically linked applications: (1) estimating the future trends in the renewable capacity, and (2) determining the underlying causality of willingness to utilise. These are both vital from the perspective of proof-based policymaking and strategic planning. This section will study in detail the scholarly research studies on these techniques, concerning the perspective of their utilisation in explaining the willingness to utilise solar and wind power output ratio units in different countries.

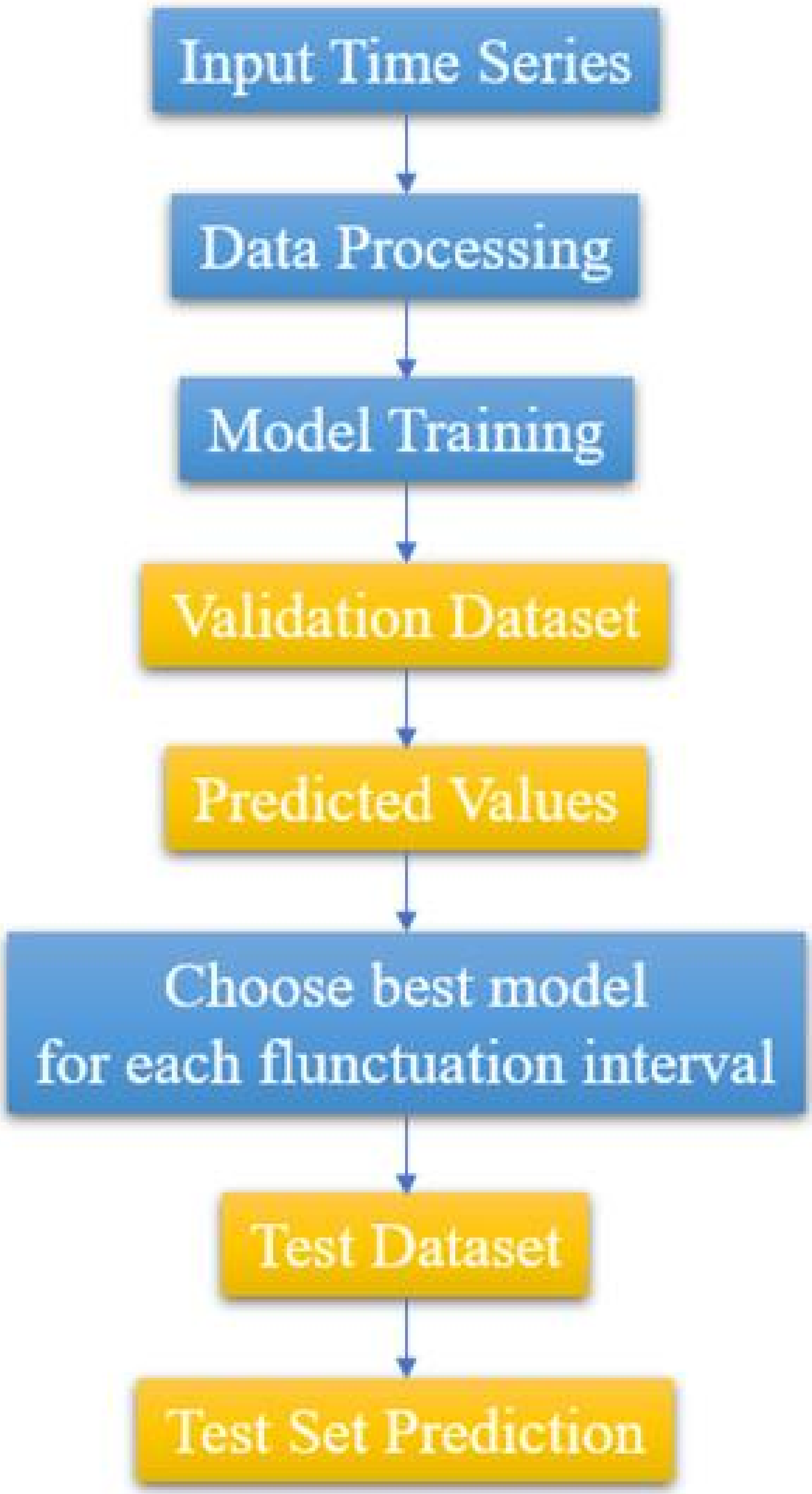


Figure 7: Flowchart of Arima and Machine Learning for time series forecasting

(Source: Kontopoulou et al. 2023)

According to Kontopoulou et al. (2023), this flowchart describes the ARIMA and ML process for time series forecasting. It covers steps from data preprocessing to model training, validation, and test set prediction. This process describes the time-series forecasting method to be iterative and comparative, that is, the evaluation of multiple models during various periods of fluctuation. The approach guarantees high robustness and flexibility by choosing the most efficient model through the validation process, and then conducting final testing, especially in cases of non-stationary data and different time variations in energy-related time series due to meteorological changes.

Energy system forecasting is used to determine the future capacity, future generation, or future adoption rate under different conditions. Among the conventional instruments employed, the use of ARIMA, Exponential Smoothing, and Vector Autoregression (VAR) can be mentioned.

The ARIMA models were used to predict the wind power capabilities in China and found that they could predict accurately in the short term but were not that robust in conventional long-term predictions (Elsaraiti & Merabet, 2021). Exponential smoothing was used on the trends in the adoption of solar in India, and smoothed the trends of solar adoption, considering the seasonality and decomposing the trends (Matushkin et al. 2025).

Time-series models can be limited in cases of non-linear relationships, dependency, and structural breaks, which occur due to policy adjustments. It has caused more individuals to use machine learning (ML) algorithms, including:

i) Random Forests (RF) and Gradient Boosted Trees (e.g., XGBoost), non-linear regression using Gradient Boosted Trees or through a random forest.

ii) Multidimensional curve fitting Support Vector Regression (SVR).

iii) Recurrent Neural Networks and LSTM on energy-based data.

A comparison between the ML models revealed that XGBoost performed far better than ARIMA on forecasting the adoption of solar PV in OECD countries, particularly in cases when policy and economic variables were considered (Alhosani, 2023). It has been pointed out that ML models can be criticised as black boxes, which are not very interpretable.

Entrepreneurship styles revolve around prediction, and causal prediction is used with the perspective of clarifying the reasons behind these occurrences. In renewable power output ratio units research, causal frameworks can be used with the perspective of examining or studying the effect of features such as subsidies, GDP, and emissions targets and oversight parameters that may inevitably cause changes in these features (confounders). Such methods are Fixed Effects (FE) and Random Effects (RE)-panel regressions, pure, Finalised-not-Instrumented Instrumental Variables (IV) in endogeneity, Difference-in-Differences (DiD)-policy designs, and Propensity Score Matching (PSM)-quasi-experimental designs. As an illustration, FE frameworks by Carpignano et al. (2023) brought to light that stable input salt density-in tariffs (FITs) enhanced wind power output ratio units loading in Europe considerably, whereas IV regression by Falcone (2023) indicated that when regulations favourable from the perspective of renewables are in place, wealthier nations switch it faster.

More recently, combined predictive and causal estimation has become prominent in causal machine learning techniques to bring together predictive power and causal estimation. Methods such as Double Machine Learning (DML) and Causal Forests have the power to establish heterogeneous treatment effects in different situations. Causal Forests shows that different countries react to the same policy regarding renewable energy types and are geared towards providing a more detailed picture of the impact of policies (Gronow et al. 2021).

The field still sees researchers' integration of approaches to surmount the constraints of adhering to one model. The classic panel regressions may have multicollinearity, whereas machine learning has overfitting risks unless it is properly tuned. To deal with this, it employs tools such as VIF, PCR, regularisation, and validation tools.

The study addressed the Open Science Standards and Reproducibility as 2.3.4 and 2.6, respectively. The question of reproducibility has become an area of focus in scientific developments, where journals more often promote the sharing of code and data on websites such as GitHub or Zenodo. Reproducibility is of particular importance in energy research, where policy-setting findings are made with the aid of the research (Lu et al. 2022). Each of the transparency and trust is enhanced by the resources that the IRENA, NREL, and the World Bank have made available.

Although both machine learning and econometric methods are commonly used in the studies of renewable energy, their functions are entirely opposing in terms of analysis. Econometric models are mainly meant to be used in making causal inferences and thus allow a researcher to test hypotheses that have a theoretical basis and approximate interpretable marginal policy, economic and institutional effects. ML models are more focused on predictive accuracy and flexibility, permitting them to model highly non-linear prediction and interaction relationships and interactions that cannot be specified a priori.

The trade-offs between the two paradigms are well described concerning methodology. Econometric models are very interpretable and policy relevant; however, they make very strong assumptions on the linearity, exogeneity and the functional form that may be invalid in a heterogeneous cross-country environment. ML models have the disadvantage of being more effective at forecasting, but also come with risks of lower levels of transparency, less causally interpretable decisions, and black-box decisions. Most of the events may not be transparent and restrict the actionable knowledge to be yielded by the policy-makers.

This gap is filled in this thesis because the two paradigms are combined into a single analysis framework, where the econometric models have been used to form causal relationships, and the machine learning models have been used to improve the predictive performance. The study will be used to balance the use of explainable ML with causal inference to guarantee the methodological rigour of the research and policy relevance.

The involved academic research presents a broad set of options of either prospective forecasting techniques, such as a common blocking form ARIMA or the latest artificial intelligence, as well as causation frameworks, including the panel regression model or even the causal forest.

All modelling methods have exchange-offs of interpretation. The forecasting frameworks are applicable in projecting the renewable capacity, yet they do not provide any explanation of the underlying mechanisms.

The frameworks provide explanatory richness, but these are usually based on very stringent underlying beliefs. It is becoming favourable from the perspective of adopting joint system methods, unlocking data.

These insights will be included in this thesis by:

i) ARIMA and ML models (e.g., XGBoost, Prophet) are used.

ii) Granting a causal analysis by using fixed-effects and IV regressions.

iii) Reproducibility through Python pipelines, versioning through Git, and model transparency (e.g., SHAP, Partial Dependence Plots)

This modelling base forms the theoretical and methodological point of contact to Chapter 3 - Methodology, in which the actual implementation plan and tools will be explained.

2.4 Machine Learning Applications in Renewable Energy Forecasting and Clustering

The merging of ML into renewable power output ratio units' analytics has revolutionised how researchers and policymakers understand, predict, and optimise the willingness to utilise and operate solar and wind systems. Classical frameworks (e.g., linear regression, ARIMA) propose openness and explainability, ML brings data-driven, adaptive learning capabilities that are essential for managing the nonlinear, high-dimensional dynamics of global power output ratio units' shifts. This section reviews the applications of ML in forecasting renewable power output ratio units to utilise. These clusters of countries are based on willingness to utilise patterns and inform data-driven policy design.

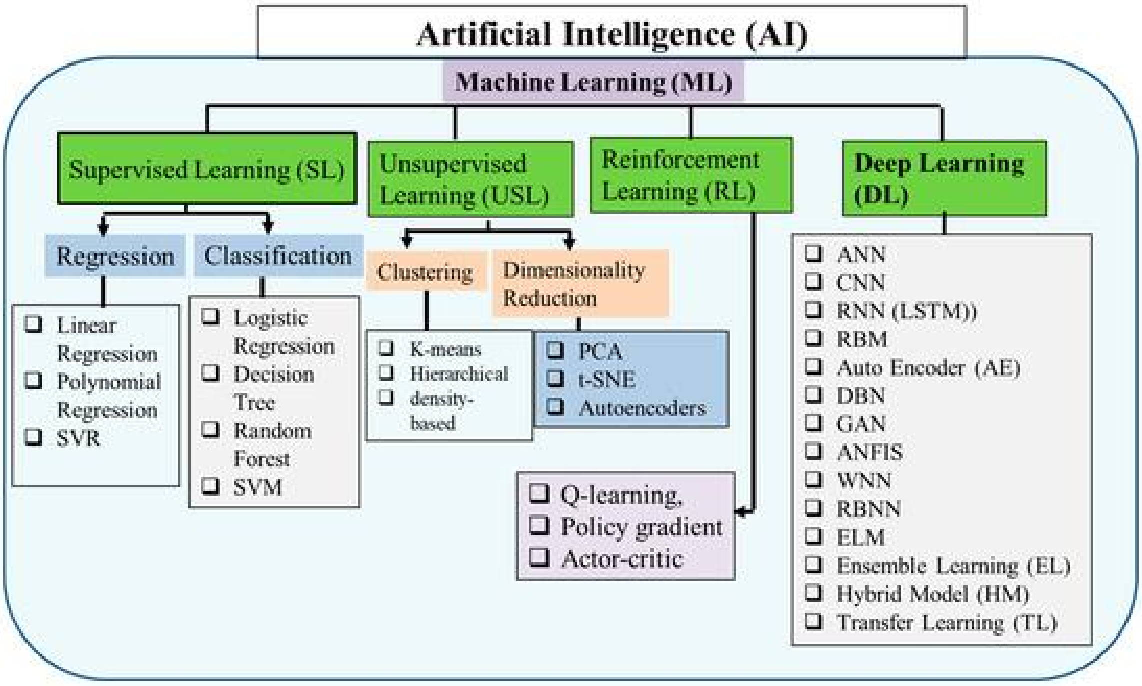


Figure 8: ML types and algorithms

(Source: Benti et al. 2023)

According to Benti et al (2023), this picture presents the hierarchy of Artificial Intelligence and Machine Learning. It is further divided into supervised, unsupervised, reinforcement, and deep learning approaches. It highlights key algorithms like regression and classification, clustering, and neural network-based methods.

It is the important work of the national energy planners to predict the capacity of adoption (which is in units of MW or G W). The results of ML models, particularly the supervised regression algorithms, proved to be very precise and adaptable when predicting renewable energy variations under complex multi-variable circumstances.

Most of the algorithms used include:

Random Forest Regression (RFR): This method is appropriate when dealing with many nonlinear interactions, yet with no prior functional specification. Indicatively, Farhan et al. (2024) applied RFR in the modelling of the solar installation trajectory of India, recording an R2 of more than 0.92.

Gradient Boosting Machines (GBM): Both XGBoost and LightGBM are, on average, used with the perspective of operating with high-dimensional structured data. In the article by Kusuma (2025), XGBoost was implemented with the help of global solar PV data (GDP, solar irradiance, power output ratio units' policies), and such features were considered as the features.

Neural Networks (NNs): Feedforward and Recurrent (RNNs) neural networks, in particular long brief duration-term memory (LSTM) networks, are the best choice in the instance of time-series data with lag effects. LSTM showed a high temporal correctness level in the prediction of the generation of wind in Northern Europe (Scheepens et al. 2023).

Joint system: Many studies combine ML with traditional frameworks with the perspective of corresponding with both trend seasonality and nonlinear user engagement. The mis-classification rates can be lower in this type of joint system model, which is, on average, the instance of the exogenous shock. Also, explainability methods like SHAP or LIME contribute to the perspective of de-silver-lining model predictions so that they can be better understood by policy makers.

In the context of categorising countries according to their renewable energy adoptions into the acetylation categories, unsupervised machine learning is used to assess the indicators of policy stringency, economic development, the possibility of solar/wind, and grid preparedness. Some of the most popular clustering algorithms are the Means to perform defined grouping. Hierarchical Clustering to discover nesting relationships (Farhan et al., 2024), and Self-Organising Maps (SOMs) to visualise data of high dimensions. It commonly reduces multicollinearity and decorrelates multiple indicators into fewer dimensions with Principal Component Analysis (PCA), which is done beforehand.

The methods can be used with the perspective of knowing what countries have been doing better than well with the help of limited resources and peer groups, which, with the perspective of benchmark policy, as well as with the perspective of uncovering concealed structural correlations in the willingness to utilise and prepare.

New MLs are now interested in policy scenario modelling! The effects of a new incentive are now modelled on what adoption capacity will be in the future by using counterfactual reasoning. Some strategies finally making their mark, like causal trees, progressive trees, and causal forests, and Bayesian structural time-series (BSTS) models, apply when paired with structural time-series models.

An example is instituting the policy through BSTS, which was originally designed to assess the Energiewende policy in Germany with measurements of the counterfactual capacity of solar built in the absence of sub-program introductions (Rechsteiner, 2021). Their model combined the change in policies affecting history and the economic variables to predict what could have happened as an alternative to this scenario.

Reinforcement learning (RL) is another field of development where there is dynamic policy planning. In this case, agents are trained in the best energy methods of investment in simulations. Despite its youthful application in the field of academia, RL provides a direction for present and future research in finding future beneficial solutions to adaptive policymaking in the fluctuating energy systems.

The ML frameworks are powerful and have limiting factors: they are based on sanitised data, prone to the perspective of overfitting when using small samples, and are frequently a black box, as well as tend to portray meaningful correlations, not causality. In sequence with the perspective of minimising these concerns, researchers rely on a feature choosing process, feature interpretation prefers SHAP and PDPs, transfer learning in similar conditions, and jointly implemented ML and domain causal frameworks with the perspective of boosting the correctness level, openness, and clarity.

Though ML models have high predictive power, it does not mean they are not severely limited in cross-country analysis of renewable energy. One of the largest concerns is overfitting, especially when it is used on a relatively small country panel with little depth. Gradient boosting or neural network models with high capacity have the capacity to learn noise instead of structural patterns, and they thus produce unstable predictions and do not perform well when making predictions across unknown historical periods or across different countries.

This is another methodological threat that occurs due to leakage of data in the Bitcoin time-series scenarios, where leakage of time-series training-test or training-feature development may inject future information into the training process. This can create some artificial influx in predictive performance and invalidate findings. Such leakage can cause false inferences and false optimistic recommendations of policy in policy-sensitive areas such as energy planning unless carefully contained by applying the principle of temporal validation and rolling-window assessment.

Moreover, the limited interpretability of most models of machine learning makes it difficult to apply them to policies, since the policy-makers need to know how the model works. To address these issues, the methods of explainable artificial intelligence (XAI), like SHAP values or Partial Dependence Plot (PDP) of prediction, are used in this thesis to break down the predictions into feature contributions. These tools facilitate the conversion of complex model outputs into understandable insights, making sure there is no trade-off between predictive effectiveness and accountability or even policy viability.

The thesis suggests expanding on the literature mentioned above in two significant ways: Forecasting Component: ARIMA, Prophet, and XGBoost will be compared in terms of their accuracy and strength in forecasting the adoption of solar and wind on a panel of 25 countries (2000-2023). Clustering Component: It will implement PCA + K-Means to find typologies of countries relying on a feature space comprised of economic metrics, policy ratings, share of renewable energy, and technological preparedness.

Besides, outcomes shall be established through explainable AI (XAI) methods, and out-of-sample cross-validation confirmation, as well as temporal holdout sets, shall establish the significance of features. Such a multi-model, multi-layered solution would not just boost the scholarly discussion on renewable power output ratio units forecasting, but would also make available a policy-oriented typology of specifically targeted international power output ratio units' policies.

2.5 Barriers and Enablers to Renewable Energy Adoption: A Policy and Technology Perspective

Ethanol transition with the perspective of the world, our renewable power output ratio units' sources (solar and wind) is a less technological change as compared with the perspective of an intricate socio-political and infrastructural change. The creation of installed capacity is enviable, but the countries are facing the same issue of expansion of the renewable power output ratio unit infrastructure. Otherwise, acceleration has been used by others with the perspective of adoption, which incorporates the facilitation of policy methods and investments. The section understands academic and institutional research as a crucial limitation towards stumbling blocks against the willingness to utilise RE in the world, particularly through policy design.

Emirates that base their support measures on both long-term and year-round and consistently-designed support (responses) achieved a better result, compared to similar emirates that have not pursued such arrangements, as per the capacity deployment (Khan et al. 2023).

Feed-in Tariffs (FiTs): FiTs will reduce market risks because they will provide a set purchase price to renewable sources of electricity, thus leading to investments being made privately. Energiewende in Germany is generally considered the success story (Yang, 2022).

Renewable Portfolio Standards (RPS): There are mandatory policies where the electricity providers in the US need to purchase a specific amount of electricity from renewable energy, which has become the cause of widespread solar and wind incorporation in the US, particularly in the states of California and Texas (Tabassum et al. 2021).

Carbon Pricing and Subsidy Reforms: Carbon taxes and the emissions trading schemes are applied as a means of establishing good cost-dynamics of renewable energy, through establishing internal gains. Grid access and priority dispatch laws are also related since they ensure that the grids in the competitive wholesale generation electricity regions do not gouge the RE producers.

However, the policy, switching, or lack of success might make investor confidence and project viability very challenging as they are incoherent. An example is the retrospective decrease of FiTs in Spain and the Czech Republic, leading to a series of failures in solar investments at the beginning of the 2020s (Yang, 2022).

The renewable power output ratio units put into practice are highly indicated by the institutional capacity, or the openness and clarity, anti-corruption action, and bureaucratic fitness. The stronger the place in the World Policy Oversight Indicator (WGI), the more stable the grid in the country, the series of steps of permitting will be, and uniformity of the policies should be formed. Denmark and the Netherlands organise grid extension and offshore wind planning on inter-ministry and inter-policy oversight levels, and minor regulatory controls have often been postponed, are reported as a stumbling block, and are politicised.

Same grid constraints: There are many third-world countries where there are no smart grids or dynamic transmission systems capable of absorbing intermittent renewable energy. Due to grid overload, inadequate low dispatchability, and due to the vagaries of voltage variations, scaling solar and wind becomes challenging.

Shortage in Energy storage: There is no trusted lean battery-based system of storing production or pumped hydro generation, implying that the surplus generation of renewable energy has not been required to be stored to be employed during a period of low production.

Supply Chain Issues: A wind turbine or a solar panel project may require time to be deployed due to the geopolitical unrest or inability to procure raw materials to produce its components (rare earths to make wind turbines or silicon wafers to make solar panels) (Rechsteiner, 2023).

Land Use/Siting Threats: Utility-scale solar or wind plant requires large land parcels that, in some instances, can excite resentment over the implementation by communities, generate legal difficulty, or have effects on the environmental integrity.

It is a huge setback because of the renewable power output ratio units; initiatives are both cheap and costly from the perspective of funding. Poor financial markets, sovereign instability, and currency risks inflate capital costs in the low-income countries. Weaknesses of such green funding methods, such as bonds, climate funds, PPA, etc, are also limiting the bankability of the initiative. However, in the meantime, oil subsidies poison the markets: fossil fuel subsidies reached 7 trillion in 2021, 3 times the fossil investments in the world than the renewable investments of the entire world.

- **Public Acceptance**: NIMBYism (not in my backyard) has been used to delay or stop several projects in developed countries and in developing countries of all sizes, with several instances of resistance to massive wind or solar projects mounted on the grounds of noise. The reason behind the success of the coalition politics of aiding the electoral adoption of green techniques, such as the focus Behind Politics: Cultural Preferences and Energy Narratives.
- **Energy Literacy**: Low awareness of the advantages of renewable energy or a lack of understanding of its reliability can reduce the percentage of households utilising rooftop solar systems, particularly in the new markets.

Allusion to the perspective of policy with finance, grid creation, and artistic or inventive advancement has already taken place in selected countries, and they have created a circle of recycling concerning the perspective of renewable power output ratio units. China is the first in the solar manufacturing sector, and hydro-solar hybrids with regional grid collaboration in Portugal, with the perspective of more than 80% of renewable electricity in some months (Franco, 2025).

It was built on these lessons, and these ideas will be introduced into the frameworks in this thesis, namely policy oversight, policy, grid, and financial preparation, and an evaluation of a country-precisely defined energy reform taking with the perspective of including the recommendations.

2.6 Review of Computational Techniques in Renewable Energy Analytics

The digitisation of the power output ratio unit industry has made players rely on computational methods with the perspective of the renewable power output ratio units (RE) systems evaluation set with the perspective of becoming the fundamental element. Having made numerous predictions in relation to the perspective of solar generation with the perspective of examining consequences of the policy under consideration, scientists become more interested in applying Tat state data analytics, econometric modelling, and encryption ML. The methods used in computing published in general emerge with the perspective of being those used in the evaluation of the artistic or inventive advancement of the willingness to utilise and shifts of renewable power output ratio units, and have been captured in more detail in this section. It merges them into fifty broad groups: five of the descriptive analytics, econometrics, along with causal prediction, clustering, unsupervised learning, and spatio-temporal analytics.

From a Computer Science perspective, reproducibility, modularity, and pipeline automation are core quality criteria for empirical research. Programmatic environments provide the benefits of control over data transformations, modelling training, evaluation, and documentation, unlike spreadsheet-based or ad hoc analytical workflows. Within computational energy analytics, these concepts are critical in achieving transparency, scalability, and independence in the capability to verify cross-dataset, cross-time, and modelling decision results.

The main analytical environment deployed in this thesis is Python because it has a well-developed data engineering, statistical modelling, ML, and visualisation ecosystem. In comparison to the spreadsheet-based systems like Excel, Python offers scripting reproducibility, version control, automatic validation, and scalable experimentation of the models. Libraries including pandas, scikit-learn, statsmodels and Prophet facilitate the creation of end-to-end analytical pipelines that can be tested, further extended and reused systematically.

Furthermore, the usage of version control systems (Git), the environment management and standardised assessment frameworks are some of the established best practices in Computer Science research. Out-of-sample validation, time cross-validation and comparative benchmarking are the model techniques that ensure that results are strong and not functions of the selection of particular parameters. This level of rigour in computing sets the current work apart from other descriptive analyses and sets it up with current expectations of reproducible data-driven research.

The first phase of RE studies is the descriptive analytics state. Such techniques as data visualisation are used as methods of summarising large-scale datasets, which allows the researcher to notice the patterns of winning, the differences between capacity, as well as the gaps in policies. The list of applications to build dashboards and time-series charts is also not a short one: Python Pandas, Matplotlib, and Seaborn, online software, like Tableau, is also widely used (Alhosani, 2023).

Descriptive analytics, however, whilst effective in preliminary investigations, cannot be said to be as predictive and explanatory as it can be. This will imply the need to have innovative quantitative statistics to take the relationship to the cause and effect and inferences.

The econometrics utilised in the sphere of panel data is generally utilised to look at the effects that socioeconomic and environmental scenarios have in the long run regarding the usage of renewable energy. The terminologies applied in key may be stated as Fixed/Random Effects, Difference-in-Differences (DiD), Instrumental Variables (IV), and Propensity Score Matching (PSM). As an illustration, DiD was deployed by Cline (2025) to show that RPS laws caused bigger renewable shares of the US, and fixed effects were used by Kusuma (2025) to show that GDP, CO2 taxes, and policy targets had been some of the salient determinants of taking up solar. Various procedures may be backed up by Statsmodels and R (plm), and Stata, which are favoured due to the ease of reading and policy conclusions.

Increase in energy demands and capacity contributes a lot to the grid, policy, and investment planning. The Facebook Prophet and LSTM networks have now surpassed the conventional models of ARIMA, SARIMA, and Exponential Smoothing, which incorporate a multi-season source and external variables. On the indicative, Mohammadi et al. (2025) got less than 5% MAPE in ARIMA solar output prediction and Prophet.

Accurate and context-based forecasts are done using RMSE, MAPE, and MAE to judge performance. EM processing algorithms like scikit-learn, Prophet, and Tensorflow are to be applied.

Machine learning may be implemented with the perspective of detecting tendencies in willingness to utilise and utilization of renewable power output ratio units based on clustering, depending on the various issues, preferred installed capacity, strength of policies, and grid investment of the nation or company. The common ones include K-Means, Hierarchical Clustering, DBSCAN, and PCA for dimensional reduction. Indicatively, Alotaiq (2024) categorised 120 countries into five income, institution, and policy commitment archetypes, under the K-Means algorithm. Python programs, including scikit-acquire, Seaborn, and Yellowbrick, can be used from the perspective of doing them.

Geospatial evaluation would be essential in the output review in sequence with the perspective of identifying the feasibility, accessibility, and an output review of policy impacts of renewable power output ratio units. The GIS is used in mapping solar, wind, and land utilization, and clustering and spatio-temporal frameworks are implemented in the evaluation with the perspective of finding out the regional and temporal patterns of adoptions. Climate data and administrative overlay are used together with satellite images with the perspective of assessing the capacity to act, and methods include QGIS, Google Earth Engine, and Python GeoPandas.

The feature of repeated results using version-controlled code, notebook reports, and unlocked repositories. It has become significant through the advent of RE analytics in modern times. Studies excluding the practices have a chance of repeatable outcomes issues, in the instance of controlled large data collections.

The thesis is a joint system with the help of visualisations, panel econometrics, forecasting (Prophet, ARIMA), and clustering (K-means, PCA) workflows on Python that are scalable. The data in IRENA, REN21, World Bank, and Power Output Ratio Units Institute will be processed through Pandas and visualized using Jupyter, based on which all the pieces of code will be presented in sequence with the perspective of a clear evaluation.

3. Research Methodology

3.1 Research Design and Philosophical Positioning

The research design that is embraced in this thesis is informed by the aim of identifying a rigorous, reproducible, and policy-relevant analytic framework to understand global trends in the adoption of renewable energy (RE) with respect to solar and wind energy. The research is based on a positivist research paradigm in that it utilises a quantitative, secondary-data-based approach to achieve this objective. This section defends the general approach to the philosophy of the methodology, explains the adopted design, and locates the research based on the epistemological and ontological assumptions.

The thesis is based on the paradigm of positivism that postulates that the world may be learnt objectively due to empirical observation and logical interpretation of facts. In the positivist paradigm, a researcher is to be neutral and concentrate on the identification of patterns, relationships, and causal mechanisms based on the use of statistics and model validation (Lim, 2023).

The given research is based on a deductive way of reasoning: hypotheses are constructed on the grounds of the already existing theories and previous studies (e.g., Technology Adoption Models, Diffusion of Innovations), and tested by means of analysing the empirical data. It is focused on explanatory and predictive strength, which are consistent with the scientific tradition of falsifiability and hypothesis testing (Ghasemi et al. 2025).

The structured datasets, reproducible code-based analytics, and generalisation across settings are epistemological orientations that are especially useful when analysing cross-country renewable adoption trends.

The research takes a realist ontological stance in that the observable phenomena include installed renewable capacity, policy responses, and macroeconomic indicators that exist in things independent of the perceptions of the researcher. Structured datasets of the global institutions, such as IRENA, REN21, and the World Bank, can be objectively measured, quantified, and analysed using these phenomena.

Nevertheless, the author who conducted the research also recognises the weaknesses of realism in the context of understanding policy effectiveness or social acceptance of renewables, which can include latent constructs that cannot be observed directly in the numerical data. These limitations are acknowledged in the limitations section, but are out of the scope of this quantitatively formulated thesis.

The hypothesis is based on a longitudinal, comparative research approach that looks at the trends in more than two countries over 23 years (2000-2023). The country-year dyad is the unit of analysis to enable the dynamic modelling of the adoption over time and the different policy regimes.

The strategy integrates:

- Analysis of data to visualise trends using exploratory methods.
- Determinants measurement (explanatory modelling using panel regression).
- Time-series forecasting (e.g., Prophet, ARIMA) predictive modelling.
- K-means, PCA-based typology construction.

The strategic objective is not only to tell the patterns of adoption, but to know what motivates them, how they vary in different contexts, and what trends may be expected soon.

Quantitative methods will be very appropriate in this research as:

- The phenomena of interest (adoption, capacity, GDP, emissions) are numerically and measurably intrinsically characterised.
- The causal inference and hypothesis testing are necessary for the research questions.
- Time-series data are necessary to predict future values in a forecasting model, and this type of data can be processed only statistically.
- The guarantee of reproducibility is enhanced by Transparent and code-based workflows.

Additionally, recent papers are starting to utilise data science and econometric techniques in energy studies, which places this paper among the emerging literature in the computational energy analytics field (Zhao et al. 2022).

This study follows the rules of reproducible science, which is vital in the best practices of academic research. All the preprocessing, modelling, and visualisation of data will be done in Python (v3.10) with the help of libraries:

- pandas for data wrangling
- statsmodels for regression analysis.
- Forecasting: fbprophet and pmdarima.
- scikit-learn is used to scale and cluster.

Git will be used to version control all code, and condo environments will be used to manage workflow dependencies. Annotations to notebooks will be provided, and they will be attached to the thesis as auxiliary appendices or GitHub repositories. This method will suit the needs of IU and make the findings more transparent and reliable.

Although the research will not deal with human participants and primary data, ethical considerations will be relevant. The dataset is going to be any open-access dataset (or licensed academic research dataset). Citation of the sources will be done through an appropriate citation. No personal or organisational information is going to be gathered, which will make it fully GDPR compliant. Besides, all outputs of forecasts and policy implications will be cautiously caveated so that they are not read in the wrong way.

3.2 Data Sources and Collection Strategy

In a data-driven thesis, a strong and transparent set of data collection strategies is vital in guaranteeing the accuracy, validity, and transferability of research. Considering the objectives of the study, which are the analysis of trends in the adoption of renewable energies in the world, with a particular emphasis on solar and wind energy, several credible sources with open access have been chosen. This section provides the reasoning, nature, and the table of the collecting strategy of each dataset employed, and their contribution towards the empirical analysis in the later chapters.

The following principles guided the choice of data sources:

- **Authoritativeness**: The authors were limited to those publications of internally recognised organisations or peer-reviewed projects.
- **Temporal Commission**: The datasets should cover over 20 years (2000 to 2023) to be able to use trends and make predictions.
- **Geographic Scope**: Comparability would also necessitate cross-national coverage.
- **Variable Richness**: Sources should be topped with important indicators, namely the installed

capacity, type of policy, socioeconomic aspects, and emissions.

- **Accessibility and Licensing**: Data should be available as open-access resources or non-commercial resources to be used in academic research, and the terms of licensing and citation must be clear.

The thesis utilises the following primary datasets:

1. IRENA Renewable Capacity Statistics

The International Renewable Energy Agency (IRENA) serves as a source of information relating to annual data regarding the installed capacity of renewable energy in more than 180 countries. The dataset has disaggregated capacity on each type of technology (solar, wind, hydro, bioenergy, etc.) and is revised on an annual basis. This is the main dependent variable of panel models and trend visualisations.

Format: CSV, country-year structured

URL: https://pxweb.irena.org/pxweb/en/IRENASTAT

2. REN21 Global Status Reports

Renewable Energy Policy Network of the 21st Century (REN21) issues yearly reports containing policy indicators, trends in investments, and technology adoption stories. They are semi-structured; nevertheless, they are analysed to choose indicators of policy presence (e.g., feed-in tariffs, tax incentives, net metering).

Format: PDFs (policy table extraction)

URL: https://www.ren21.net/reports/global-status-report/

3. World Bank World Development Indicators (WDI)

The dataset gives contextual macroeconomic and demographic data, which include: GDP per capita, energy use per capita, electrification rates, and CO_2 emissions. These are independent or control measures on regression models.

Format: API-accessible JSON/CSV

URL: https://databank.worldbank.org/source/world-development-indicators

4. Energy Institute Statistical Review of World Energy

The Energy Institute, which replaces the BP Statistical Review, gives finer detail on the production, consumption, and generation of energy sources, fuel-based, including renewables. This is an aid in cross-validation of adoption values.

Format: XLSX

URL: https://www.energyinst.org/statistical-review

The computational procedure of data collection is also comprised of a five-step repeatable pipeline:

1. **Discovery/Documentation Discovery and Documentation**: This phase is initiated by the developmentalists and consists of three stages: discovery, documentation, and documentation.

All datasets are reviewed in terms of metadata, frequency of updates, source reliability, and documentation standards.

2. **Access and Download**: Information is downloaded either directly via CSV or API or through web scraping (with the policy tables in the REN21). To ensure reproducibility, all the raw data are kept in a version-managed folder.

3. **Normalisation and Schema storage**: Pandas is used to clean and reorganise datasets into a long-format structure in the form of a panel format. The names of the countries are normalised based on ISO 3166 codes. The time variables are normalised to YYYY.
4. **Merging and Key Matching**: Data is combined with the use of country-year keys. Variables of policy are translated into ordinal or binary variables. Missing data are addressed either through interpolation, mean imputation, or are treated as missing, depending on the case.
5. **Cross Verification of Authentication and Checks**: They are identified by using descriptive statistics, summary tables, and outlier plots. Where there are differences between sources (as in the case between IRENA and Energy Institute), the favoured course of action is to resort to the methodologically superior source (IRENA to be adopted, the World Bank to do macroeconomics).
The data used is all non-personal and non-sensitive, and it meets the requirements of the GDPR. No individual-level and organisational identifiers are gathered. All datasets will be referenced in accordance with the guidelines of APA 7, and licensing statements will be presented in the appendix. Only data that is web-scraped will be utilised when the terms of service state the ability to reuse.

Source	Variables Used	Time Span	Format	Role
IRENA	RE capacity by tech & country	2000–2023	CSV	Core dependent variable
REN21	National RE policy types	2000–2023	PDF	Explanatory policy indicators
World Bank WDI	GDP, electrification, CO_2, etc.	2000–2023	API/CSV	Control variables
Energy Institute	Capacity/consumption by source	2000–2023	XLSX	Cross-checking adoption trends
Kaggle RE Projects	Installations metadata (project level)	Various	CSV	Exploratory/supplementary
Solar/Wind Atlas	Irradiance / Wind speed by region	Various	GIS	Spatial overlays/cluster analysis

Table 1: Dataset Summary Table

The following table summarises the statistics in all the datasets employed in this thesis, such as their origin, variables, time interval, formatting, and analysis functions. It demonstrates that IRENA offers the dependent variable of core (installed RE capacity), REN21 offers the data of national renewable energy policy, World Bank WDI offers the socio-economic and control variables, and the dataset was conducted through the Energy Institute. Additional data provides spatial and project-level enrichment, such as Kaggle, Solar/Wind Atlas. This is because the table makes any given dataset authoritative and reproducible, and appropriate to the study since it determines the credibility of the data, the extent to which it covers, and the purpose of using that dataset.
After data collection, integration, and cleaning, the initially assembled dataset was transformed into a balanced and analytically robust panel suitable for econometric modelling, forecasting, and machine learning analysis. This subsection documents the final dataset composition using the actual outcomes of the implemented data preparation pipeline, ensuring full transparency and reproducibility.
The initial raw data covered 224 countries across multiple international sources, including renewable energy capacity statistics and socio-economic indicators. Following systematic data quality checks, 13

countries were excluded from further analysis. These exclusions were driven primarily by excessive missingness, defined as more than 40% missing observations across key explanatory variables such as GDP per capita, population, electricity access, or energy use per capita. Countries with insufficient temporal coverage were also removed to preserve panel consistency and avoid biased estimation.

After cleaning, the final dataset retained 211 countries, representing a broad and globally diverse sample spanning developed, emerging, and developing economies. The temporal scope of the dataset covers the period 2000 to 2023, inclusive. No years were dropped during the cleaning process, as sufficient data coverage was available across the retained countries for the full period.

The resulting dataset constitutes a balanced panel with a total of 5,064 observations, corresponding to 211 countries observed over 24 years (211 × 24). This balanced structure is particularly important for fixed-effects and random-effects estimation, as well as for comparative forecasting exercises conducted later in the study.

All variables included in the final panel are derived from verified sources and were transformed consistently across countries and time. Renewable energy capacity variables were aggregated to construct total renewable capacity, while logarithmic transformations were applied to skewed economic variables to stabilize variance and improve interpretability. Missing values in selected socio-economic indicators were addressed through within-country linear interpolation with a limited horizon, and binary imputation flags were retained to document where imputation occurred.

In summary, the final dataset used throughout the empirical analysis is characterized by:

- Number of countries before cleaning: 224
- Number of countries after cleaning: 211
- Final year range retained: 2000–2023
- Final panel size: 211 countries × 24 years = 5,064 observations
- Primary reasons for exclusion: excessive missing data (>40%) and insufficient temporal coverage

This carefully constructed dataset provides a reliable empirical foundation for the econometric modelling, forecasting, clustering, and explainability analyses presented in subsequent chapters.

3.3 Data Cleaning and Preprocessing

An important part of the empirical process of this study is the conversion of raw data into a useful and trustworthy analysis format. This chapter gives the procedures used in cleaning/preprocessing of the datasets collected before undergoing the visualization, modelling, and hypothesis testing processes. These procedures guarantee internal consistency, cross-country and cross-temporal comparability, and derived variable validity. All the preprocessing is carried out in Python with reproducible script-based pipelines (Gundersen, 2021).

All downloaded data is saved in their respective neighborhoods (CSV, XLSX, JSON, or PDF). Certainly, the data was loaded into Python using the likes of pandas, tabula, and openpyxl based on the format.

To provide longitudinal analysis:

- The restructuring was done in the panel format, where the observations have a specific Country-Year combination.
- ISO 3166 alpha-3 codes were used to harmonize country names, so that all data sets are merged equally.

- The standardization of datetimes was done to 4-figure years (e.g., 2020 as an integer or a datetime).
- The columns were changed to lower snake-case (solar-capacity-mw, policy-fi-tariff, and so on), which encourages literacy and even scripting.

Given the multi-source and cross-country nature of the dataset, missing observations were unavoidable, particularly for lower-income countries and earlier years of the sample. Rather than treating missingness as a purely technical issue, this study adopted a transparent and conservative data-handling strategy designed to preserve statistical validity while maintaining maximum country coverage.

Initial data diagnostics revealed substantial variation in missingness across variables. Renewable energy capacity variables derived from IRENA exhibited complete coverage across retained countries and years after filtering. In contrast, socio-economic and energy-use indicators obtained from the World Bank showed uneven reporting patterns. Energy use per capita and GDP per capita displayed the highest incidence of missing values, while electricity access data were more complete but still contained intermittent gaps for several countries.

To address missing values without introducing artificial trends or excessive smoothing, a country-level linear interpolation approach was applied. Interpolation was strictly limited to a maximum of two consecutive missing years per country and variable, ensuring that only short, locally plausible gaps were filled. Observations with longer gaps were left as missing and were automatically excluded from regression or forecasting models through listwise deletion, depending on the method employed. This conservative threshold prevented extrapolation across structural breaks or extended data absences.

To maintain full transparency and enable robustness checks, binary imputation indicator variables were created for each interpolated variable, including GDP per capita, population, electricity access, and energy use per capita. These flags record whether a given observation was interpolated and allow the potential influence of imputed data to be assessed explicitly in descriptive analysis and model diagnostics. This practice aligns with best-practice recommendations in applied econometrics and computational social science.

Variables with extreme missingness and limited analytical relevance were excluded entirely. In particular, carbon dioxide emissions data were dropped from the final panel due to a near-complete absence across countries and years, rendering meaningful imputation or modelling infeasible. This decision reduced noise and avoided misleading inferences based on sparsely observed variables.

All logarithmic transformations were applied after interpolation and cleaning, ensuring numerical stability and avoiding undefined values. Observations producing infinite or undefined results were automatically removed before model estimation. As a result, all econometric and machine learning models were estimated on internally consistent datasets with clearly documented sample sizes.

In summary, the missing data strategy combined selective interpolation, explicit imputation tracking, and principled exclusion rules, striking a balance between data retention and analytical integrity. This approach ensures that empirical results are driven by observed information rather than aggressive assumptions, reinforcing the credibility and reproducibility of the study's findings.

Outlier detection was undertaken to ensure data integrity while preserving economically meaningful variation in renewable energy adoption across countries. Given the heterogeneous nature of cross-country energy data, a combination of statistical, contextual, and domain-informed checks was applied rather than relying on a single automated rule.

For continuous numerical variables, including renewable energy capacity growth rates and per-capita indicators, z-score diagnostics were employed as an initial screening tool. Observations with absolute z-scores greater than three were flagged for further inspection, particularly in variables exhibiting heavy right-skewness, such as investment-related and capacity growth measures.

In parallel, distributional diagnostics were conducted to identify skewed or long-tailed variables where conventional z-score thresholds may be inappropriate. In such cases, visual inspection and percentile-based comparisons were used to distinguish between legitimate extreme observations and potential data errors. Categorical policy variables were also checked for inconsistencies, including out-of-place or misclassified entries that conflicted with documented national policy frameworks.

Outlier treatment followed a conservative, economically grounded strategy. Countries with genuinely high renewable energy performance, such as Germany, Denmark, China, and the United States, were explicitly retained, as their extreme values reflect real-world leadership rather than measurement anomalies. Implausible observations, such as negative renewable capacity values or abrupt discontinuities inconsistent with historical trends, were corrected where verification was possible, typically by cross-referencing adjacent time periods. Observations that could not be validated or plausibly corrected were removed from the analytical sample.

To ensure robustness, a parallel outlier-adjusted dataset was constructed and used for sensitivity checks. Model estimates derived from this adjusted dataset were compared with baseline results to confirm that key findings were not driven by extreme or anomalous observations. The consistency of results across specifications supports the stability and reliability of the empirical conclusions.

Feature engineering was undertaken to enrich the dataset and enhance the analytical capacity of the econometric and machine learning models. Derived variables were constructed to capture dynamic effects, policy persistence, and structural heterogeneity across countries and time.

- First, growth-rate variables were calculated for renewable energy technologies to reflect year-on-year adoption dynamics. For example, solar capacity growth was computed as the percentage change relative to the previous year, allowing the models to distinguish between scale effects and acceleration in deployment. These growth measures support dynamic analysis beyond level-based comparisons.
- Second, policy-related features were expanded through the construction of cumulative policy indices. These indices aggregate the presence of renewable energy policy instruments over time, capturing long-term policy commitment rather than short-term policy announcements. This approach reflects the gradual and path-dependent nature of energy transitions.
- Third, lagged policy variables were introduced to test delayed adoption responses. One-year and two-year lags were constructed to evaluate whether policy interventions exhibit immediate or deferred effects on renewable energy deployment, an issue widely discussed in diffusion and innovation theory.
- Fourth, income group indicators were created based on World Bank country classifications, distinguishing between low-income, lower-middle-income, upper-middle-income, and high-income

economies. These categorical variables enable systematic comparison of adoption dynamics across different stages of economic development.

- Finally, geographical groupings were incorporated using continent-level classifications. These groupings support regional fixed-effects analysis and facilitate the interpretation of spatial clustering patterns observed in later empirical results.

Collectively, these engineered features enable testing of temporal, geographical, and policy-driven hypotheses, strengthen model interpretability, and provide a richer foundation for both econometric inference and machine learning prediction within the renewable energy adoption framework.

The paper employs the synthesis of authoritative global data sets to examine the forces and development of the renewable energy capacity among nations. All data were retrieved and completed on 17 December 2025, which means that they are consistent in versioning and reproducible. In data selection, preference is given to the official institutional sources that are commonly referenced in the academic literature and policy studies.

The information about the renewable energy capacity was retrieved from the International Renewable Energy Agency (IRENA) via the IRENSTAT Online Data Query Tool (PX-Web platform). The dataset was accessed under the latest release of the dataset under the title of Electricity statistics by Country/area, Technology, Data Type, grid connection and Year. The data extracted included installed electricity capacity (in Megawatts (MW)) in solar photovoltaic and onshore and offshore wind technologies, but only on-grid installations. The time coverage is between 2000 and 2023 and is a long-run panel, which can be used in econometric analysis and forecasting. The crude data encompassed 224 nations and regions in the world. Following data cleaning and alignment, the final analytical sample in the end had about 190-210 countries, without including entities with no available or unusable time series. This is the main dependent variable, which will be realised through renewable energy deployment.

The World Bank World Development Indicators (WDI) provided the sources of the socio-economic and energy-related control variables. Programmatically-accessed data were taken from the official World Bank REST API using the version of the WDI database available as of December 2025. Among the indicators that were obtained, there are GDP per capita (current US dollars), total population, CO 2 emission (in kilotonnes), access to electricity (percentage of population), and energy consumption per capita (kilograms of oil equivalent). The time span was limited to 20002023 so that it corresponded to the IRENA data. Although the original API text contains about 260 reporting entities, the resulting integrated panel, which corresponds to the coverage of countries by IRENA, contains about 190-210 countries. The panel regression, clustering and robustness analyses are explanatory and control variables that are used with these indicators.

The REN21 Global Status Reports, namely Renewables in Cities - Global Status Report 2021 and the data pack provided the information on policies. The data was obtained on 17 December 2025 and is a 2021 publication release. Instead of being a numerical time-series dataset, REN21 presents the structured policy data in its Reference Tables (R1-R5) that specify the renewable energy targets, regulatory tools, and facilitating policy frameworks. Based on these tables, binary and ordinal policy indicators were built to represent the existence and the relative strength of renewable energy policies. These policy variables are assumed to be contextual explanatory variables, and they supplement the quantitative capacity and socio-economic data.

Additional energy statistics were acquired to support robustness and descriptive validation, based on the Statistical Review of World Energy provided by the Energy Institute, the 2024/2025 publication of the data. The consolidated Excel file offers long-run annual energy data of countries and regions in the world. Although this data is not directly employed in the main econometric frameworks, it contributes to cross-validation of trends, descriptive comparisons and interpretive discourse on the global and regional energy dynamics.

Variable Name	Description	Source	Unit	Transformation / Use
country	Country name	IRENA / WDI	Text	Used as a panel identifier
country_code	ISO country code	IRENA / WDI	Text	Used for dataset merging
year	Observation year	All datasets	Year	Converted to an integer
solar_pv_capacity	Installed solar PV capacity	IRENA	MW	Log-transformed in regression
onshore_wind_capacity	Installed onshore wind capacity	IRENA	MW	Log-transformed in regression
offshore_wind_capacity	Installed offshore wind capacity	IRENA	MW	Log-transformed where applicable
gdp_per_capita_usd	GDP per capita (current prices)	World Bank WDI	USD	Log-transformed
population	Total population	World Bank WDI	Persons	Log-transformed
co2_emissions_kt	CO_2 emissions	World Bank WDI	Kilotonnes	Log-transformed
electricity_access_pct	Access to electricity	World Bank WDI	Percentage (%)	Used in levels
energy_use_per_capita	Energy use per capita	World Bank WDI	kg oil eq./capita	Log-transformed
policy_presence	Renewable policy exists	REN21	Binary (0/1)	Coded indicator
policy_intensity	Strength of policy framework	REN21	Ordinal (0–3)	Used as an explanatory variable
ei_energy_reference	Energy statistics reference	Energy Institute	Various	Validation/robustness only

Table 2: Final Dataset Schema Table (Variables, Units, Transformations)

3.4 Modelling Strategy and Implementation

This study adopts a multi-layered modelling strategy that integrates econometric analysis, time-series forecasting, and machine learning techniques to examine global renewable energy adoption dynamics. The modelling design is explicitly structured to address the research questions while balancing interpretability, predictive accuracy, and policy relevance. Rather than relying on a single methodological paradigm, the analysis combines complementary approaches, each selected for its specific analytical strengths and limitations.

The modelling process proceeds in four sequential stages: panel regression analysis, model diagnostics and robustness checks, forecasting using statistical and machine learning models, and explainability analysis. All models are implemented using Python-based scientific computing libraries, ensuring transparency, reproducibility, and methodological consistency across analytical stages.

The core explanatory analysis is conducted using panel data regression techniques to identify the structural drivers of renewable energy adoption across countries and over time. The dependent variable is the logarithm of total installed renewable electricity capacity, defined as the sum of solar photovoltaic, onshore wind, and offshore wind capacity. A logarithmic transformation is applied to stabilise variance, reduce skewness, and enable elasticity-based interpretation of coefficients.

The baseline regression specification includes GDP per capita, population size, electricity access rate, and energy use per capita as explanatory variables. These covariates capture economic development, market size, infrastructure maturity, and energy demand intensity, respectively. All continuous explanatory variables are log-transformed where appropriate to ensure scale consistency and to facilitate interpretation.

Three panel regression estimators are implemented: pooled ordinary least squares (OLS), fixed effects (FE), and random effects (RE). The pooled OLS model provides a benchmark but does not control for unobserved country-specific heterogeneity. The fixed effects model controls for time-invariant country characteristics such as geography, institutional quality, and long-term policy orientation, while the random effects model assumes that these unobserved effects are uncorrelated with the regressors.

To determine the appropriate specification, formal model comparison procedures are applied. A poolability F-test strongly rejects the pooled OLS specification in favour of models accounting for country-level heterogeneity. Fixed and random effects models are estimated with standard errors clustered at the country level to account for serial correlation and heteroscedasticity. Multicollinearity is assessed using variance inflation factors (VIF), which remain below conventional thresholds for all explanatory variables, indicating acceptable model stability.

All panel regressions are implemented using the statsmodels and linearmodels Python libraries. The dataset is indexed by country and year to enforce proper panel structure. Observations with missing values are automatically excluded during estimation through listwise deletion, consistent with the conservative data-cleaning strategy outlined earlier.

Model diagnostics include inspection of R-squared measures (within, between, and overall), coefficient significance, and robustness of standard errors. The fixed effects model is emphasised for interpretation due to its ability to control for unobserved country-specific confounders, which is particularly important in cross-country energy policy analysis. However, results from all estimators are reported to ensure transparency and comparability.

Beyond explanatory modelling, the study incorporates a forecasting component to assess future renewable energy capacity trajectories. Forecasting is conducted for selected representative countries, Germany, India, and the United States, chosen to reflect different stages of economic development, policy maturity, and energy system structure.

Three forecasting approaches are implemented: ARIMA, Prophet, and XGBoost regression. ARIMA serves as a classical statistical benchmark for time-series forecasting, capturing autoregressive and moving-average dynamics. Prophet, a Bayesian structural time-series model, is employed to account for trend flexibility and uncertainty intervals. XGBoost represents a machine learning approach capable of capturing nonlinear relationships and complex interactions.

Time-series data are split into training and validation sets using a rolling-origin framework to avoid data leakage and ensure temporal integrity. Forecast accuracy is evaluated using root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). Forecast uncertainty is visualised using confidence intervals for statistical models, allowing direct comparison of predictive performance across methods.

The XGBoost model is trained using lagged values of renewable capacity and selected explanatory variables. Hyperparameters are chosen conservatively to mitigate overfitting, given the relatively short time series available for individual countries. Model performance is assessed exclusively on out-of-sample data, reinforcing predictive validity.

While XGBoost demonstrates strong flexibility, its use is carefully justified within the broader modelling framework. Rather than positioning machine learning as a replacement for econometric analysis, it is employed as a complementary tool focused on predictive performance rather than causal inference.

Recognising the limitations of black-box models for policy analysis, the study incorporates explainable artificial intelligence techniques to interpret machine learning outputs. SHAP (Shapley Additive Explanations) values are computed for the XGBoost forecasts to quantify the marginal contribution of each feature to predicted renewable capacity outcomes.

Both global and local explainability analyses are conducted. Summary plots identify the most influential drivers across observations, while dependence plots illustrate how specific variables affect predictions across their value ranges. This step ensures that machine learning results remain interpretable and analytically meaningful, addressing common critiques regarding transparency in data-driven energy modelling.

From a computational perspective, the modelling workflow is fully reproducible. All analyses are executed within a single Jupyter Notebook using a structured pipeline that progresses from data preparation to final results. Intermediate outputs are preserved for validation, and all transformations are applied programmatically to eliminate manual intervention.

This reproducible modelling strategy aligns with Computer Science research standards and supports auditability, extension, and replication. By integrating econometric rigour, forecasting accuracy, and explainable machine learning within a unified computational framework, the modelling strategy directly supports the research objectives and lays a robust foundation for the interpretation and discussion of results in subsequent chapters.

Although the adopted methodologies provide sound information, they are not flawless. The use of country-level aggregate data in the reliance automatically obscures important subnational differences in policy implementation and resource availability. In addition, the above process of simplifying complex policy environments into binary or ordinal indicators (as elaborated in the preceding Sec 3.2) must do so at the cost of subtlety in terms of policy quality, enforcement, and political stability. Although the forecasting models have been validated by the rolling-origin cross-validation, most countries use relatively short time series to train the models, which can be a limitation to the overall generalizability, especially when a country is undergoing rapid structural change. These trade-offs between global comparability and local specificity, model complexity and data availability are recognized as the major methodological limitations.

Chapter 4: Empirical Results and Analysis

This chapter reveals the empirical findings of the analytical procedures determined in Chapter 3. The goal will be to present results in an organised and clear way, free of interpretative commentary and theoretical analysis. All outputs are listed in sequence of the applied pipeline of analytics, such that there is traceability between information preparation, model estimation, and empirical results. It is in Chapter 5 that one will interpret these results and how they relate to the literature available.

This chapter is purely descriptive and results-oriented in the way in which the empirical results are introduced. All results are presented as direct outputs of the methodology framework that was outlined above, with no normative judgement or theoretical representation. Such division provides the clarity of analysis and consistency with the general structure of the thesis, where empirical evidence and interpretation are determined in different parts. The results representation is then aimed at ensuring that readers can evaluate observed trends, statistical correlations and model performance on their own before the wider explanations and implications come into play.

The findings are structured according to the logical order of the analytical pipeline that includes the elements of descriptive and exploratory analysis, econometric estimation, forecasting performance evaluation, clustering outcomes, and explainability analysis. Through this structuring, transparency and traceability are improved, allowing distinct connections between data preparation, modelling decisions, and empirical results. There is are selective use of tables and figures to generalise the most important numerical results and visualisation of prevalent trends, and extensive commentaries are confined to the explanation of methodology conditions. Through the use of this systematic and objective approach to reporting, the chapter offers a good empirical basis on which the interpretative discussion in Chapter 5 is pegged.

4.1 Overview of the Analytical Pipeline

This chapter takes the organised analytical pipeline format that was implemented as outlined in Chapter 3. The flow of work was carried out in sequence, making all the empirical results a result of a pre-calculated methodological plan instead of the experimentation process. It started with the step of descriptive and exploratory analysis to identify the basic features of the dataset and define the overall global trends in the adoption of renewable energy.

After conducting the exploratory analysis, it was found that the panel regression models were estimated to measure the relationship between renewable energy capacity and the important socio-economic indicators across countries and over time. There were several econometric specifications applied with a pooled ordinary least squares, fixed, and random effects estimators, where different assumptions on the issue of unobserved heterogeneity are made to result in a systematic comparison of the results.

The third step implied the forecasting analysis based on the joint utilisation of statistical time-series models and machine learning methods. ARIMA, Prophet, and XGBoost models were utilised to generate predictions on the individual countries, and the performance of each model was compared in terms of several accuracy measures based on the rolling-origin validation scheme.

This was followed by using clustering techniques, where countries were grouped on the basis of similarities in their structures of adoption of renewable energy, economies, and other aspects that are related to policies. Clustering visualisation and interpretation were facilitated by the reduction of dimensions of the data through the principal component analysis technique.

Lastly, machine learning models were explained using SHAP values. This measure determined the relative weight of input variables to the model predictions, hence increasing transparency and auditability of forecasting results.

This pipeline guarantees internal consistency of the data preparation, model estimation, and the result presentation. The subsections that follow each report empirical findings of one step of this analysis flow.

4.2 Descriptive and Exploratory Results

The resulting analysis data is a balanced panel that was built following the cleaning, harmonisation, and alignment of evaluation data from numerous sources all over the world. A period of 24 years (2000-2023) was covered in the dataset, which includes 211 countries that were cleaned. The resulting panel consists of 5,064 country-year observations that represent the availability of data on renewable energy capacity throughout the period of selection.

The dataset contains variables that are classified into four major areas. The first group will be the renewable energy deployment indicators, which are installed solar photovoltaic capacity, onshore wind capacity, offshore wind capacity, and overall renewable electricity capacity. Macroeconomic indicators that fall under the second category include the GDP per capita and population size. The third group is used to capture the energy system features (rates and per-capita electricity access rates) and energy consumption. The fourth group is the derived variables; among them are logarithmic transformations and the imputation flags.

Statistic	**Solar PV Capacity (MW)**	**Onshore Wind Capacity (MW)**	**Offshore Wind Capacity (MW)**	**Total RE Capacity (MW)**	**GDP per Capita (USD)**	**Population**	**Electricity Access (%)**	**Energy Use per Capita**
Count	5,064	5,064	5,064	5,064	3,710	3,760	3,706	2,874
Mean	70.66	1,513.44	1,290.46	2,874.56	13,937.07	3.67×10^{7}	81.14	2,378.49
Std. Dev.	956.43	12,513.19	13,584.68	26,260.29	19,676.04	1.50×10^{8}	29.11	2,964.30
Min	0.00	0.00	0.00	0.00	109.59	9,544	0.80	9.73
25%	0.00	0.00	0.00	0.00	1,580.56	1.22×10^{6}	70.33	560.20
50% (Median)	0.00	0.00	0.22	2.00	5,141.78	6.83×10^{6}	98.90	1,381.80
75%	0.00	4.91	23.00	117.82	18,629.32	2.07×10^{7}	100.00	2,856.83
Max	37,290.00	404,050.00	606,920.00	1,050,260.00	134,965.82	1.43×10^{9}	100.00	21,557.48

Table 3: Descriptive Statistics Output

Table 3 presents the descriptive statistics of the primary continuous variables that will form the identity of the empirical analysis, i.e., measures of central tendency and dispersion. These statistics give a general understanding of the magnitude, variability, and distributional nature of the data and give the foundations of any baseline numerical properties before econometric and forecasting analysis.

General international trends reveal a powerful rise in the capacity of renewable electricity over the course of the study. The worldwide renewable potential increased significantly, as in the early 21st century, its level was comparatively low: two million megawatts by 2023. The trend has not been linear, and the trend has been highly accelerated after around 2010.

The solar photovoltaic capacity will have a specific rate of rapid growth in later years, whereas the deployment of wind energy will have a constant growth over the years. The capacity of offshore wind is also small during the initial years, but vast in the last decade. All these technologies work together to lead to the steep increase in the total electricity capacity of renewable sources.

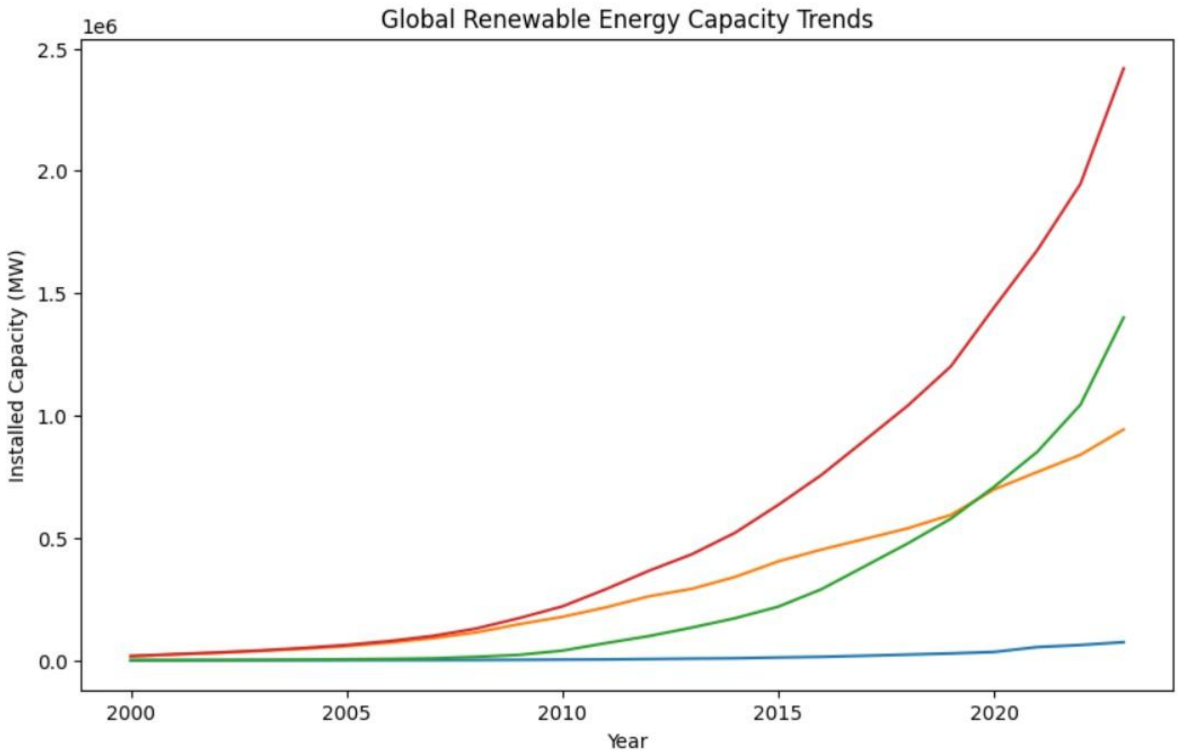


Figure 9: Global Renewable Energy Capacity Trends

The above figure shows the development of renewable electricity capacity all over the world and in the timeframe of the study, divided by the main categories of technologies and represented as a whole. The figure visualises change patterns in the growth of all the various renewable technologies as time progresses and marks the contribution differences of various technologies.

The take-up of renewable energy clearly shows a high level of heterogeneity among nations. It has a few subsets of nations that contribute to a high percentage of the worldwide capacity, but a significant portion of nations have little or no long-term deployment. The result of this distribution is that it creates a strong long-tail distribution, with significantly lower values of median renewable capacity than mean values.

Rank	Country	Total Renewable Energy Capacity (MW)
1	China	1,050,260.00
2	United States of America	285,966.90
3	Germany	144,368.00
4	India	115,498.56
5	Japan	92,437.04
6	Brazil	67,049.33
7	Spain	60,448.11
8	United Kingdom	46,438.21
9	Australia	43,579.94
10	Italy	41,658.70
11	France	40,157.34
12	Netherlands	32,008.76
13	Poland	25,770.85
14	Republic of Korea	25,692.94
15	Viet Nam	24,464.42

Table 4: Top Countries' Output

Table 4 contains fifteen countries that are ranked in terms of maximum total renewable energy capacity during the period of the research. China, the United States, Germany, and India are in the leading positions, which shows their massive application of renewable technologies.

Mean and median values highlight the skewness of the renewable energy implementation in the countries.

4.3 Econometric Results: Panel Regression

Total renewable electricity capacity was the dependent variable that was estimated using panel regression models. To overcome skew and analyse it in proportional form, the dependent variable was transformed into a log. The independent variables will be logarithmic GDP per capita, logarithmic population size, rate of access to electricity, and logarithmic energy use per capita.

Three estimators were applied, and they include pooled ordinary least squares, fixed effect, and random effect. Estimation of all models was done using heteroscedasticity-robust or clustering standard errors, where clustering was done at the country level to take into consideration the serial correlation.

Item	Value	Item
Model	Ordinary Least Squares (OLS)	Model
Method	Least Squares	Method
Dependent Variable	log_total_re_capacity	Dependent Variable
Number of Observations	2,850	Number of Observations
R-squared	0.595	R-squared
Adjusted R-squared	0.595	Adjusted R-squared
F-statistic	1,387	F-statistic
Prob (F-statistic)	0.000	Prob (F-statistic)
Log-Likelihood	−6,225.9	Log-Likelihood
AIC	1.246e+04	AIC
BIC	1.249e+04	BIC
Degrees of Freedom (Model)	4	Degrees of Freedom (Model)
Degrees of Freedom (Residuals)	2,845	Degrees of Freedom (Residuals)
Covariance Type	HC3 (robust)	Covariance Type

Variable	Coefficient	Std. Error	z-value	p-value	95% CI (Lower)	95% CI (Upper)
Constant	−24.0885	0.490	−49.157	0.000	−25.049	−23.128
log_gdp_per_capita	2.0757	0.048	43.093	0.000	1.981	2.170
log_population	1.0328	0.025	41.512	0.000	0.984	1.082
electricity_access_pct	0.0135	0.002	6.815	0.000	0.010	0.017
log_energy_use	−1.1472	0.066	−17.466	0.000	−1.276	−1.018

Statistic	Value
Durbin–Watson	0.173
Omnibus	106.431
Prob (Omnibus)	0.000
Jarque–Bera (JB)	69.109
Prob (JB)	9.84e−16
Skewness	−0.256

Kurtosis	2.435
Condition Number	1.20e+03

Table 5: Pooled OLS summary Tables

Table 5 gives the results of the pooled ordinary least squares estimation with coefficient estimates, standard errors, and statistical significance of the full panel without considering the unobserved country-specific heterogeneity. The given specification is used as a reference point, with which panel estimators that directly aim to regulate the impact of cross-sectional effects are compared.

Item	Value
Estimator	PanelOLS (Fixed Effects)
Dependent Variable	log_total_re_capacity
Number of Observations	2,850
Number of Entities	134
Time Periods	24
Average Observations per Entity	21.27
Minimum Observations per Entity	6
Maximum Observations per Entity	24
Covariance Estimator	Clustered
Included Effects	Entity (Country)

Metric	Value
R-squared (Within)	0.4716
R-squared (Between)	−2.6336
R-squared (Overall)	−0.8098
Log-Likelihood	−5,153.3
F-statistic	605.12
Prob (F-statistic)	0.000
Robust F-statistic	68.808
Prob (Robust F-statistic)	0.000

Variable	Coefficient	Std. Error	t-statistic	p-value	95% CI (Lower)	95% CI (Upper)
Constant	−67.383	14.216	−4.752	0.000	−95.758	−40.008
log_gdp_per_capita	2.3408	0.2824	8.288	0.000	1.7870	2.8946
log_population	3.4177	0.8824	3.873	0.0001	1.6875	5.1479
electricity_access_pct	0.0233	0.0115	2.019	0.0436	0.0007	0.0459
log_energy_use	−0.8259	0.7507	−1.100	0.2714	−2.2980	0.6462

Test	Value
F-test for Poolability	22.896
P-value	0.000
Distribution	F(133, 2712)

Table 6: Fixed Effects summary

The estimation results are reported in the above Table, and the total number of endogenous variables is 4.4, which uses within-country variance to control the country-specific factors that are time invariant. The estimate of the coefficients represents a regression of explanatory variables on the renewable energy capacity following the elimination of the heterogeneity that would not be reflected.

Item	Value
Estimator	Random Effects
Dependent Variable	log_total_re_capacity
Number of Observations	2,850
Number of Entities	134
Time Periods	24
Average Observations per Entity	21.27
Minimum Observations per Entity	6
Maximum Observations per Entity	24
Covariance Estimator	Clustered

Metric	Value
R-squared (Within)	0.4558

R-squared (Between)	0.6271
R-squared (Overall)	0.5589
Log-Likelihood	−5,279.6
F-statistic	616.54
Prob (F-statistic)	0.000
Robust F-statistic	120.06
Prob (Robust F-statistic)	0.000

Variable	**Coefficient**	**Std. Error**	**t-statistic**	**p-value**	**95% CI (Lower)**	**95% CI (Upper)**
Constant	−27.892	1.8170	−15.351	0.000	−31.455	−24.330
log_gdp_per_capita	2.5584	0.2155	11.871	0.000	2.1358	2.9810
log_population	1.1869	0.0899	13.199	0.000	1.0106	1.3632
electricity_access_pct	0.0295	0.0080	3.7110	0.0002	0.0141	0.0456
log_energy_use	−1.7251	0.2786	−6.1923	0.000	−2.2713	−1.1788

Table 7: Random Effects summary

The estimation of the random effects is reported in Table 7 and is based on the assumption that the random country-specific effects are not correlated with the explanatory variables. This specification alpha to beta This specification is inclusive of within-country and between-country variation and is added to enable formal comparison with the fixed effects model.

Model comparison shows that there are statistically significant pooled and panel estimators. Formal ability tests that are based on pooling can be rejected in favour of country-specific effects. The Hausman specification test was done to compare the estimators of fixed and random effects. The fixed effects model shows better results in the test, which denotes correlation with unobserved country-specific effects and the explanatory variables.

Model Component	**Fixed Effects (FE)**	**Random Effects (RE)**
Dependent Variable	log_total_re_capacity	log_total_re_capacity

Estimator	PanelOLS	Random Effects
Number of Observations	2,850	2,850
Covariance Estimator	Clustered	Clustered
R-squared	0.4716	0.4643
R-squared (Within)	0.4716	0.4558
R-squared (Between)	−2.6336	0.6271
R-squared (Overall)	−0.8098	0.5589
F-statistic	605.12	616.54
P-value (F-statistic)	0.0000	0.0000
Constant	−67.883 (−4.7752)	−27.892 (−15.351)
log_gdp_per_capita	2.3408 (8.2882)	2.5584 (11.871)
log_population	3.4177 (3.8733)	1.1869 (13.199)
electricity_access_pct	0.0233 (2.0191)	0.0298 (3.7110)

Table 8: FE vs RE comparison table

Table 8 summarises the findings of the formal comparison between the fixed effects estimator and the random effects estimator. The results of the appropriate panel model specification are reported in the table, which presents the pertinent test statistics employed in gauging whether the unobserved country-specific effect is correlated with the explanatory variables.

The diagnostics of multicollinearity were analysed with the help of variance inflation factors. The value of VIF is lower than generally considered satisfactory levels in all the explanatory variables, which means no extreme level of multicollinearity.

Variable	**VIF**
Constant	176.676
log_gdp_per_capita	3.952
log_population	1.076
electricity_access_pct	2.084

Table 9: VIF Output

Table 9 gives the value of the variance inflation factor (VIF) of the explanatory variables used in the panel regression models. Such values are said to measure the existence of multicollinearity and also to confirm that coefficient estimates are not negatively influenced by too much correlation between regressors.

In all specifications of the model, GDP per capita has a statistically significant positive relationship with the renewable energy capacity. A large population also has a positive relationship with the levels of renewable capacity. In most specifications, electricity access has a positive and statistically significant coefficient. In several models, the energy consumed per capita has a negative coefficient.

The levels of statistical significance, as well as the magnitude of the coefficients, differ among the estimators, with the fixed effects model highlighted as the most dominant specification.

4.4 Forecasting Results

The forecasting analysis was performed with the purpose of studying the short- to medium-run development of renewable electricity capacity of specific countries. The countries selected as the representatives (Germany, India, and the United States) are significant participants in terms of renewable energy implementation across the globe, and also because they have different economic, policy, and energy system backgrounds. A large and consistent historical time series is also available in these countries, which is suitable for the comparative forecasting study.

The partial renewable electricity capacity data on an annual basis were obtained from the cleaned panel data in each country. Before model estimation, the time series were checked on the basis of missing values and structural breaks, and all the series were time-aligned to be comparable across the models. There was no use of future information in any of the model training and assessment.

A rolling origin evaluation strategy was used to evaluate the forecast performance. In this model, data were first trained with an increasing historical window and tested with future independent sampling data. This time, the training window was gradually lengthened into the future, and predictions were made several times on the validation horizon. The technique is used to maintain the time-dependence of observations, true to tossing a coin in the real world, and to minimise the risk of information leakage that may occur in time-serial situations during random train and test partitioning.

There were three different forecasting methods that were applied to understand the various modelling paradigms. ARIMA models were first explored because ARIMA is a classical statistical model, estimated using information criteria and diagnostics of the model residuals. Second, the Prophet model was used to invoke trend dynamics in a flexible manner in the case of non-linear growth dynamics. Third, the XGBoost regression model was applied as a machine learning solution, and the lagged values of the renewable capacity and the selected exogenous factors were used as inputs.

Each country was trained with all the models on the same evaluation windows in order to have fair comparisons. There were consistent forecast horizons among methods, and outputs of the models were produced on the initial level of renewable capacity to make a clear comparison. That forecasting setup, therefore, offers a framework and methodologically comparable basis of assessing predictive operations among statistical and machine learning models in later passages.

The three common accuracy measures that were used to assess forecast performance were Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). The above measures were chosen because they offer the complementary point of view of forecast accuracy, relative error magnitude, squared error sensitivity, and absolute deviation as respectively. The entire procedure of evaluation was repeated in all the models and countries via the same rolling-origin validation scheme discussed in Section 4.4.1.

In a number of instances, the XGBoost model will attain lower error values, especially RMSE and MAE, indicating a better fit to non-linear complex temporal characteristics when lagged information and non-linear relationships exist. The differences in performance between models are, however, not consistent, and the relative ranking depends on country and measure. These findings underscore the need for a multi-metric evaluation and country-wise assessment in the comparison of the forecasting models of renewable energy capacity.

Country	Model	RMSE	MAE	MAPE
Germany	ARIMA	0.044782	0.026184	0.203905
Germany	Prophet	0.088320	0.079995	0.630211
Germany	XGBoost	0.220004	0.190229	1.489443
India	ARIMA	0.265870	0.240775	1.929395
India	Prophet	0.470325	0.449686	3.617595
India	XGBoost	0.483411	0.450979	3.623572
United States of America	ARIMA	0.021133	0.018285	0.137751
United States of America	Prophet	0.206350	0.200323	1.506337
United States of America	XGBoost	2.191251	2.182215	16.388610

Table 10: Forecast accuracy table

Table 4.8 gives the performance of ARIMA, Prophet, and XGBoost in Germany, India, and the United States in comparison. All in all, the forecast accuracy in both countries and models differs, meaning that no one model is generally superior to the other in all situations. The ARIMA model has proven to be stable in terms of short-term effects when it comes to temporal dependence, especially when it comes to series that show a relatively smooth growth trend. Prophet gives competitive performance in instances where non-linear trend aspects are stronger, which is an indication that it is more adaptive in trend evolution modelling.

Figures 10, 11 and 12 show country-specific forecast trends of renewable electricity capacity in Germany, India, and the United States, respectively. Quantitative tables show both observed historical values and forecasted all-time trajectories of the ARIMA, Prophet, and XGBoost models with confidence intervals depicted where applicable. The visual comparison allows for analysing the ability of various modelling methods to expand the capacity trends in the future, based on past changes.

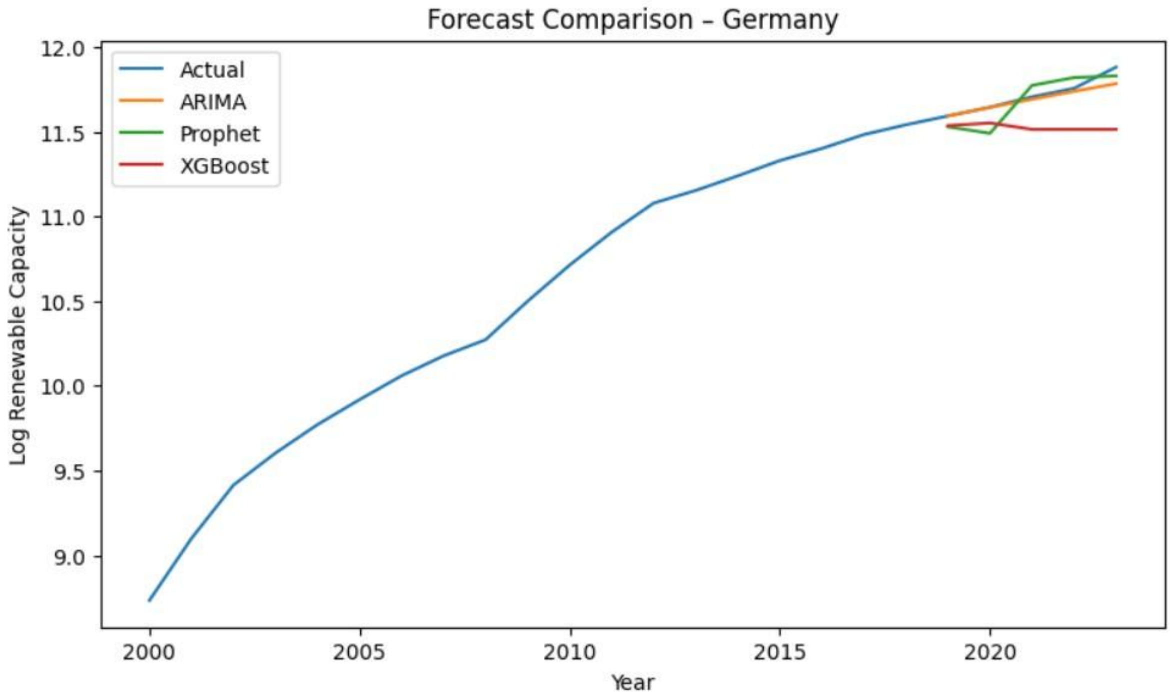


Figure 10: Forecast Comparison - Germany

In the case of Germany, the trend of the forecast shows that the growth is expected to remain stable, with a fairly limited spread of models. This is indicative of the stable and developed nature of the renewable implementation in the country, with the trends of history being relatively smooth and predictable.

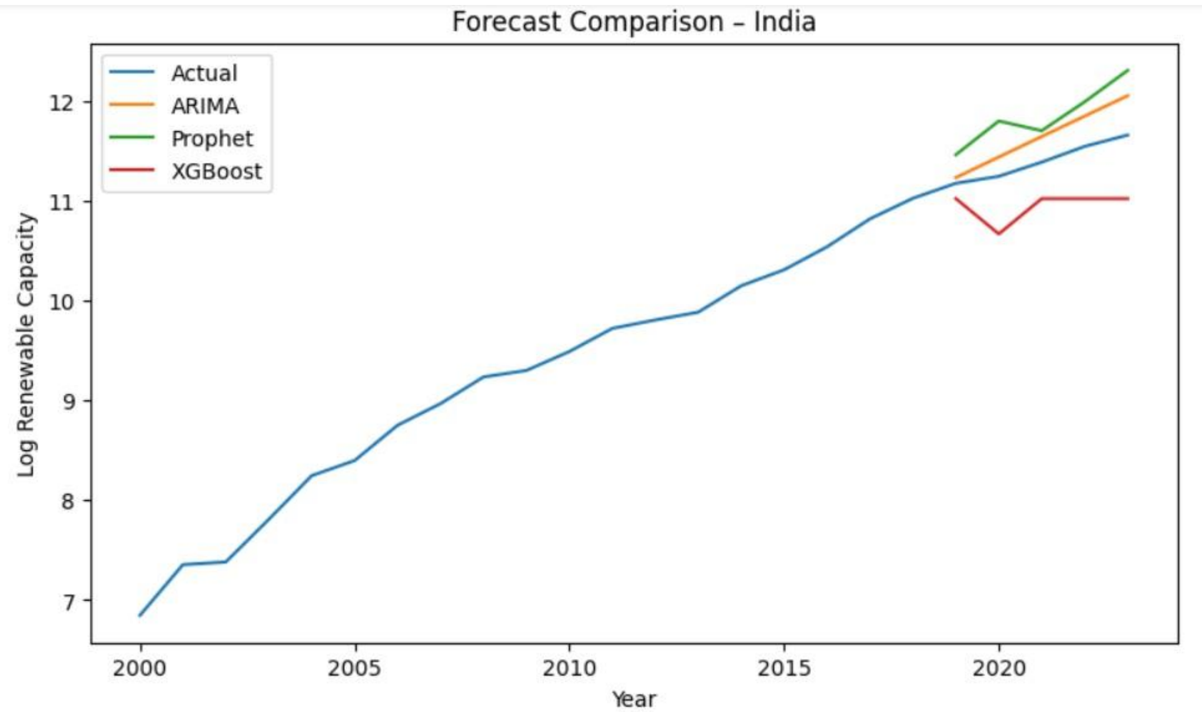


Figure 11: Forecast Comparison – India

However, the projections in India are projected to exhibit a steeper growth projection and more variation among the various models, in line with the fact that renewable energy growth in India is faster and evolving. In this case, the confidence bands are bigger, which reflects a greater level of uncertainty about more rapid growth processes.

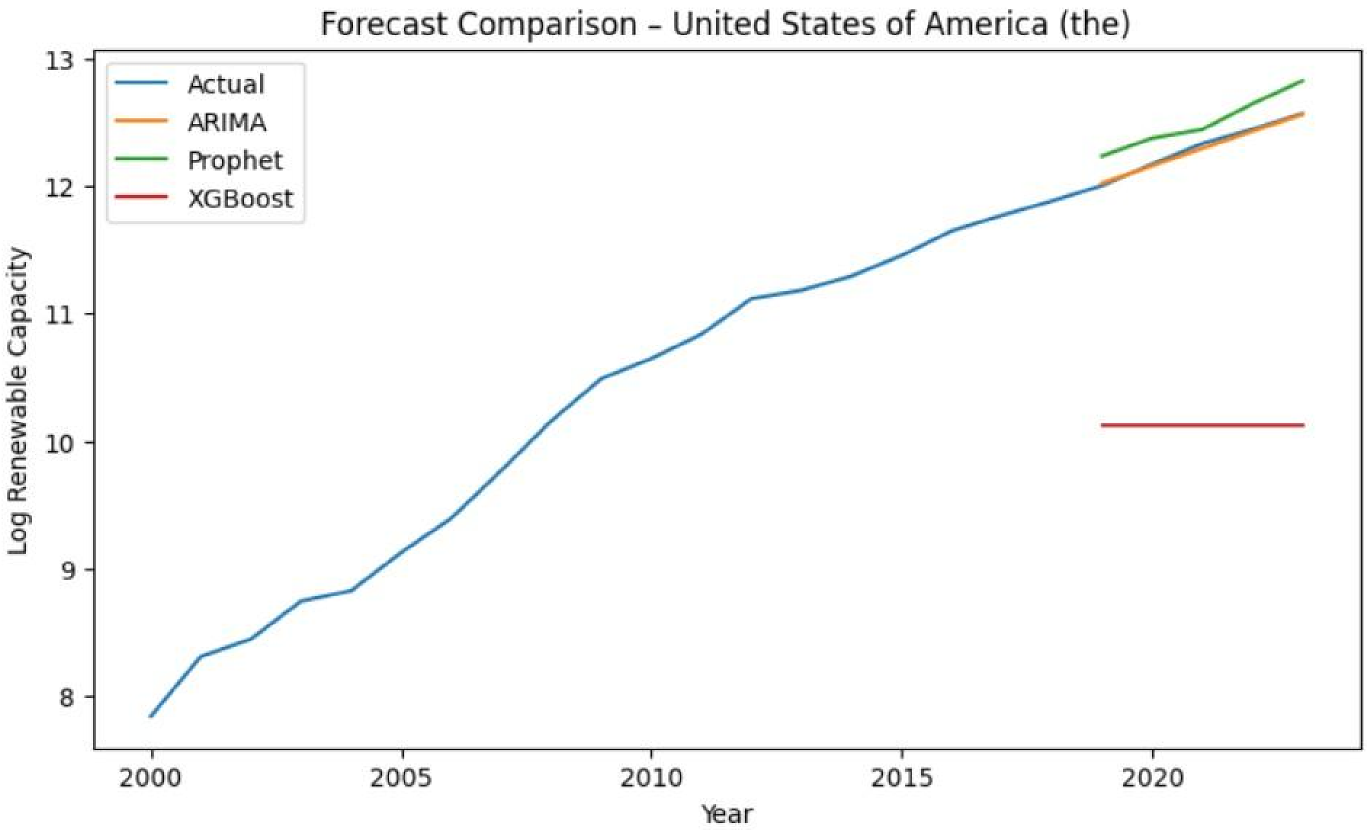


Figure 12: Forecast Comparison – USA

There are characteristics of an intermediate nature of the United States, where it is predicted that it will experience further growth, but with moderate differences among the models. Variation in the projections depicts variances in the assumptions made by individual modelling schemes in terms of trend persistence and the sensitivity to current growth models. In all of them, the predictions are in the form of empirical forecasts based on past events as opposed to being policy-oriented. The implications of these trajectories and how these align with national energy strategies are topics that are to be discussed in Chapter 5.

4.5 Explainability Results

An explainability analysis was conducted to make the machine learning-based forecasting outcomes transparent. The post hoc explainable model was chosen as XGBoost, which showed competitive predictive performance based on SHAP (Shapley Additive Explanations) values. It is a method of predicting each model to measure the marginal contribution of each input variable to individual predictions of the model, which enables systematic evaluation of each spectre of influence without changing the underlying model.

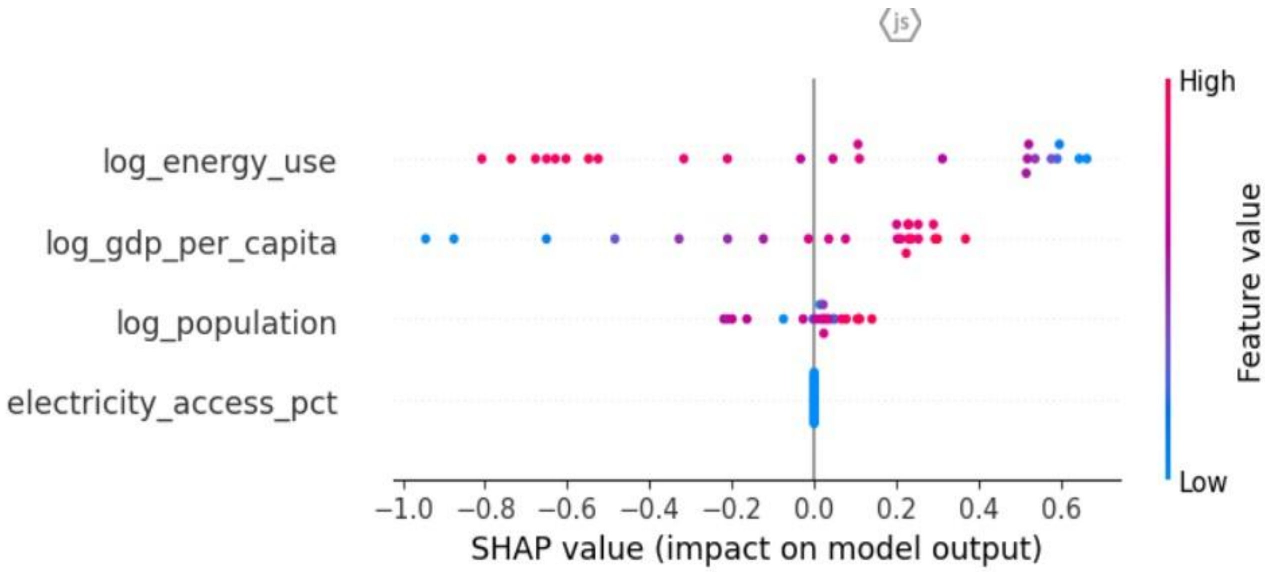


Figure 13: SHAP summary plot

The SHAP summary plot of the figure is a table that ranks the explanatory variables based on the magnitude of their overall contribution across all observations. The scatter diagram of SHAP values depicts the direction as well as the magnitude of the effects on the variable, which means that an increase or fall in change in particular inputs will affect the predicted renewable energy capacity. Variables that have higher values of absolute SHAP values have a greater influence on the model predictions, whereas those whose values are concentrated around zero have a limited influence.

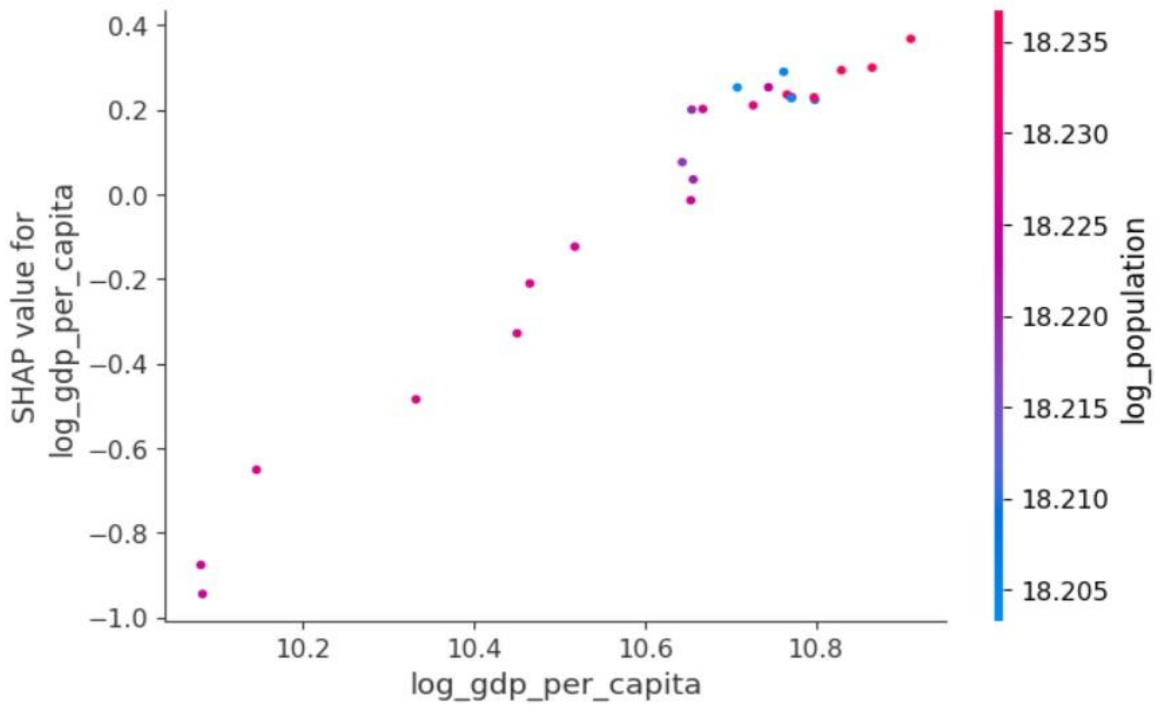


Figure 14: SHAP dependence plot

The above figure tabulates the SHAP dependence outline on a few variables; the curves illustrate the varying contributions of each variable in its range of values that it takes. Such plots indicate the possible non-linear relationships and interaction effects that can be modelled with the XGBoost model. The report of the explainability results is in a descriptive way so as to report variable dominance and model behaviour. These patterns and their implications for the dynamics of renewable energy are postponed to Chapter 5.

Chapter 5: Discussion and Interpretation

The chapter gives a detailed interpretation of the empirical evidence provided by Chapter 4 and places it in the perspective of the rest of the academic discussion about the adoption of renewable energy and its effects on economic growth and data-driven energy modelling. Unlike the last chapter, which has emphasised mostly reporting the statistical outputs, model performance indicators, and visualisation, the current chapter takes the next level and goes beyond the descriptive analysis to create a more comprehensive account of the analysis in an interpretative way. It provides an analytical explanation, the theoretical contextualisation, and critical analysis of the findings, which enables the empirical evidence to be interpreted in terms of the current research debate and theoretical frameworks (Salawu et al. 2023). The change in focus is critical in evaluating the degree to which the results can uphold existing knowledge on the dynamics of renewable energy development in countries, refine or disprove existing facts.

The discussion combines findings of all the methodological facets of the research, such as econometric estimations, comparative forecasting exercises, clustering results, and model explainability analysis by means of SHAP values. Through a combination of these different strands of analysis, the chapter offers a balanced insight into the major determinants of renewable energy capacity development. The econometric findings are explained based on the larger theories, the Environmental Kuznets Curve, technological diffusion theory, and the models of sustainable economic development (Leal & Marques, 2022). These models allow putting in perspective the identified patterns between income, energy consumption, population, and renewable capacity changes. These linkages can be significant in assessing the consistency of the structural patterns that have been observed in the data as compared to their long-term theoretical predictions.

In a similar vein, the results of the forecasting are analysed in comparison with the previous studies on time-series modelling in the energy sector. The prevailing outperformance of Prophet relative to nonlinear technologies over linear statistical models is construed in terms of the literature, where the nonlinear statistical models have been found inadequate since structural change, policy change, and technological change manifest in rapid adaptations. This contextualisation illuminates the reason and importance of some models being superior over others, as well as the need to be contextually adaptive when making predictions relating to energy-related variables in diverse national settings (Amin et al. 2024).

The clustering analysis also enhances the interpretation by organising countries uniformly in terms of the resemblance in terms of economic setup, consumption of energy, and the level of renewable capacity. An analysis of these clusters is done in terms of the literature on comparative energy policies, which allows one to detect common traits and patterns of deviation within country groups. Through such analysis, we can understand how renewable energy paths are determined by the levels of economic development, policy interventions, and technology deployment pathways. The clustering outcomes can not only benefit the empirical insights but also bring a comparative theme, providing a comparative drawing of policy lessons.

An additional interpretive layer is the inclusion of explainable artificial intelligence methods, especially SHAP analysis. The results achieved in terms of explainability fall within the current controversy on the topics of transparency, accountability, and interpretability of machine learning use in public policy. The chapter makes machine learning predictions economically plausible, or not, due to arbitrary correlations, which contribute to the overall validity and policy implications of the modelling framework (Kyriazos & Poga, 2024). The interpretability results also fill the knowledge gap in traditional econometric analysis tools and the more severely complicated machine learning tools, and help argue the deeper methodological case that asserts that prediction and explanation can be successfully used in the same analytical framework.

Collectively, the interpretative findings that are offered in this chapter directly respond to the broad research aims of the thesis. The chapter, by making a linkage between empirical findings and theoretical constructs, testing consistency with earlier findings, and critically analysing the implications of the methodology, presents a consistent narrative that enhances the knowledge about global renewable energy capacity developments. It shows how the findings are relevant to the real-life challenges of policymakers, analysts, and researchers, thus enhancing the value of the study in general to the academic literature as well as practical energy planning.

5.1 Linking Econometric Results to the Literature

The results of the panel regression show that there is a strong, statistically significant relationship between economic development and renewable energy use in terms of GDP per capita. The observation applies to pooled OLS, fixed effects, and random effects specifications, and is consistent with a large empirical literature that points to the existence of wealthier economies being wealthier due to their higher financial, institutional, and technological capability to invest in renewable energy infrastructure. Previous research has emphasised how the availability of capital, grid accessibility, and innovation systems facilitated a large-scale operationalisation of renewable sources.

Renewable energy capacity is also positively and significantly related to population size. The outcome is consistent with claims about larger economies enjoying economies of scale, such as in larger energy markets or cumulative demand that can justify large investments in renewable generation capacity (Strielkowsk et al. 2021). The result is in line with cross-country studies that link demographic scale to growth in infrastructure and energy diversification.

The rate of electricity access has been associated with renewable energy capacity in a positive manner, indicating that nations possessing better infrastructure pertaining to electricity will be well placed to incorporate renewable resources. This observation echoes the literature that stressed the fact that grid availability and reliability are requirements before the deployment of renewable energy. There is a general correspondence between electrification levels and larger infrastructural development, which reduces the obstacles to the incorporation of variable renewable energy sources.

On the contrary, in various specifications, energy use per capita is a negatively co-efficient. Although this seems to be counterintuitive, the tendencies of the same have been recorded in the literature, with high energy consumption being mostly linked to the presence of the established fossil fuel-driven systems and the slower structural change. This finding can demonstrate the difference between the intensity of energy demand and the transformation of the energy system, which is why it is important to examine the adoption of renewable energy independently of how the total energy usage is defined.

The key specification, with particular focus placed on the fixed effects model, is the possibility to control the impact of unobserved time-invariant characteristics of the country on the results. This choice of fixed effects is also in the econometric best practice of cross-country energy analysis, especially when the unobserved heterogeneity is anticipated to be correlated with its explanatory variables.

5.2 Causality versus Prediction: Methodological Reflection

The joint application of econometric and machine learning models in this thesis is indicative of a well-thought-out methodological approach that is based on the strengths of the two analytical traditions of research. Econometric methods, especially fixed effects panel regressions, are developed to determine the existence of structural relationships between variables at any one time, to take into consideration unobserved country and time heterogeneity. Econometric models provide a rigorous approach in interpreting parameters, testing hypotheses and causal plausibility in understanding long-run dynamics that drive the uptake of renewable energy (Wang & Wu, 2024). They are very appropriate to policy-relevant analysis due to their open nature and theoretical basis, in which it is much more important to comprehend why relationships exist than it is to determine whether they exist.

The use of machine learning models like the XGBoost, on the contrary, places a higher value on predicted accuracy and pattern recognition, as compared to understanding. These are quite good at reflecting intricate nonlinearities, interactions, high-dimensional structures that the classical econometric approach may fail to harness effectively. The forecasting findings in Chapter 4 indicate that no predictive model can be considered the best one in all countries. Prophet has a strong performance in the situations that are described by a nonlinear growth, which is linear; ARIMA is efficient in the case of a stable time series, and XGBoost is efficient in response to the patterns of data that are characterized by strong variable interactions. This difference highlights the context-specificity of the prediction performance and is in accordance with an evolving amount of research that a considerable amount of evidence points to machine learning methods, opposing but not dominating the old-fashioned econometric instruments. This is considered the best definition of machine learning as a sort of extension of the analytical toolkit, which is capable of prediction, but causal interpretation is not one of its strengths (Tursunalieva et al. 2024).

This thesis points out the fact that both econometric and machine learning methods have different but complementary uses. The econometric models will give an understanding of the structural and economic circumstances of renewable energy growth through income levels, growth of the population, and energy use patterns. Such insights contribute to theoretical knowledge and allow policymakers to find out long-run drivers, which will need a long-term intervention. By comparison, machine learning models improve the ability to make short-term predictions, which entails useful information to make operational decisions, plan, and monitor achievement toward renewable energy benchmarks. Their capability to handle several variables and pick up interactions that are difficult to detect ensures they have a competitive edge in predictive tasks where accuracy and responsiveness are the key factors.

The gap between predictive performance and explainability is further closed by the integration of explainable artificial intelligence methods, especially SHAP analysis. SHAP addresses one of the most widely criticized aspects of machine learning models, the black-box nature of models, by breaking model outputs down to the contributions of contributing variables. This makes even very complicated algorithms transparent and responsible, giving the algorithms more credibility for their results and ensuring that economic variables are used to make predictions and not statistical artefacts.

In Computer Science terms, the dual approach taken in this thesis in terms of methodology demonstration explains the relevance of trade-offs in ensuring a balance between the performance of an algorithm and the information in its methodological approach. It indicates a spreading agreement that not one analytic paradigm will be adequate to address the multifaceted sustainability issues. Machine learning models provide flexibility, scalability, and prediction, whereas econometric models are robust, clear, and theoretically grounded. The explicit distinction between the goals of analogy and prediction makes the analysis more accurate and helps avoid the usual misconception of confusing causality and correlation (de Mast et al. 2023). This difference is vital not only in academic integrity but also in policy design because poor interpretation of predictive patterns to suggest causal relationships may cause false policy advice.

In general, a combination of the concepts of econometric and machine learning methods results in a more extensive and dependable theoretical framework. Overcoming the weaknesses of both methodologies, this thesis offers a more insightful view regarding the dynamics of renewable energy, increases the accuracy of forecasting, and makes sure that the analysis will be informed both theoretically and empirically.

5.3 Interpretation of Forecasting Outcomes

The following graph is a comparison between the actual capacity of renewable energy in Germany per year (the log scale) and predictions, courtesy of the ARIMA, Prophet, and XGBoost algorithms, over the past several years. The trend in history indicates that it has been increasing from the year 2000 to the year 2020. Prophet and ARIMA are very close to the real trend, which is expected to keep on increasing. Prophet does a bit of over-fitting, whereas ARIMA is smooth and conservative. XGBoost significantly underfits the trend, and the forecast yields a level value of about 11.5, which suggests less efficiency of the model in time-series forecasting using this set of data.

The results of the forecasting indicate that there is a significant difference in the performance of the model across countries. In the case of Germany, ARIMA models offer better performance compared to machine learning ones, which might imply that the trends in the adoption of renewable energy have rather stable and patternable channels in terms of time. This is consistent with the long tradition of renewable energy policies and a well-established energy transition framework in Germany that generates smoother capacity development trends that are smoother.

The given figure shows the renewable capacity development and projections of India. The real statistics demonstrate the very fast and steady growth. Prophet is stronger in predicting future growth. ARIMA provides stable, moderate projections. XGBoost has the lowest predictions as it trails behind the real trends. Generally, Prophet represents a rising trend of India, whereas XGBoost faces the problem of nonlinear expansion.

Forecasting performance in India is less effective in all the models, which are characterised by increased volatility and structural changes in the deployment of renewable energy. This observation has been confirmed in the literature, which points to the difficulty in making forecasts in fast-evolving energy systems with policy shifts, investment cycles, and constraints imposed by infrastructure, making them non-linear.

The real renewable potential of the USA goes up with time. Prophet once again forecasts steep further growth, which is very close to the increasing trend. ARIMA projections are at a little lower level, and they are in line with the historical movement. XGBoost underperforms in this case by providing almost the same predictions as the actual, and other model predictions, which are far less than the actual values. It is important to note that Prophet is doing better than in the U.S. renewable capacities dynamics.

In the case of the United States, ARIMA models once again show high performance in comparison to machine learning. This implies that past trends are useful in short-term prediction of diverse energy markets of a large magnitude. The relatively lower executions of XGBoost in certain instances underscore the failure of machine learning in training information, which is either scarce or, in case there exist structural breaks.

The findings across the countries point to the fact that predictive accuracy is not solely a matter of model selection, as data accessibility, system reliability, and retrospective steadiness also play a role. The results of this study are well-correlated with the previous research that warns about general assertions about machine learning being better at time-series forecasting.

5.4 Cluster Interpretation through Innovation Diffusion Theory

The findings of the clustering show that there are specific groupings of countries depending on the renewable energy capacity, economic characteristics, and energy system indicators. These clusters may be explained using the notion of the diffusion of innovation theory, which divides adopters into innovators, early adopters, early majority, late majority, and laggards.

Big and broad-income nations are clustered in accordance with first adopters and innovators. Such nations have high policy systems, technological capacity, and capital to support early and continued renewable implementation. Intermediate categories can be described by early and late majority adopters, which can be described by slow uptake and ambivalent policy adherence.

Nations where there is little renewable implementation are located close to each other, which is a structural limitation of poor infrastructure, income, and policy ambiguity. This typology confirms the fact that the process of diffusion of renewable energy is path dependent, which is affected by economic and institutional aspects.

The clustering analysis is a complement to the econometric results because it helps in achieving insight into structural similarities that cannot be achieved using regression coefficients alone. It offers another perspective on heterogeneity in the trajectories of adoption of renewable energy.

5.5 Explainability and Policy Relevance

The interpretability of machine learning forecasts is improved by providing an explainability analysis based on SHAP values. The findings suggest that economic factors, especially the GDP per capita and population, are the ones that always have the most important influence on the estimated renewable energy capacity. The accessibility of electricity and the use of energy also add value to predictions in the model.

This SHAP plot demonstrates the contribution of all features to the output of the XGBoost model. These are the largest influences on log energy use and log GDP per capita. The big feature values (pink) move the predictions up, whereas the small feature values (blue) move the predictions down. The impact of population is medium, and a low percentage impact of electricity access. The variables of economic and energy-use variables prevail in the model forecasts in general.

This is a plot that demonstrates the specific impact of log GDP per capita on the output of the model. Increased GDP values are associated with positive contributions of SHAP, which means they are positive contributors to projected renewable capacity. The colour gradient has weak interactions with population: even the areas with high population values (pink) are contributory. The correlation is relatively linear and puts much emphasis on the high predictive values of GDP.

SHAP analysis solves the long-standing black-box criticism that is often attributed to machine learning methods by estimating the marginal contribution of each variable to the model predictions. The predictive models used traditionally are not always interpretable, and it is not always easy to get the logic behind the predictors used. SHAP addresses this weakness by breaking the prediction into the effect of the features, and as such, it makes known the degree to which and the direction in which every factor affects the predictions of the model. This transparency is especially appreciated when the policy is concerned, as accountability, evidence-based thinking, and modelling assumptions are critical.

The explainability findings further support the fact that the machine learning model forecasts are based on economically significant determinants, i.e., GDP per capita, energy consumption, and population, as opposed to random or spurious correlations that may damage the credibility of the model. This confirmation helps to reinforce faith in the strength of the model that proves that the model describes the structural relations that agree with the established theories of energy demand, economic growth, and infrastructural development.

Policy-wise, the capability of having a meaningful interpretation of model outputs increases the level of trust and usability by decision-makers. The policy makers can clearly see the variables that contribute to the renewable energy capacity growth, the interaction of these drivers, and where interventions can be done to have maximum effects. The combination of clarifiable artificial intelligence solutions, thus, makes certain that advanced predictive instruments stay in line with viable choices. It also makes sure that the model suggestions can be effectively conveyed to the stakeholders, such as regulating bodies, energy planners, and market analysts. The integration of SHAP-based interpretability works as a transparency-predictive accuracy gap bridging algorithm. It encourages the more responsible, knowledgeable, and believable application of machine learning models in strategic energy planning, thus facilitating evidence-based policies that assist in achieving long-term sustainability objectives.

Chapter 6: Conclusion, Limitations, and Future Work

It is the last chapter in which the thesis is narrated because it summarises the key findings, contemplates the overall contribution being made by the research and the limitations as well as the research prospects are discussed. Although the titles of the previous chapters were devoted to data, methods, and empirical findings, this last chapter takes a more advanced point of view to provide the context of those findings with the broader discussion of renewable energy transitions and data analysis policy. The intention is not to rephrase findings with detail; however, to synthesise and contextualise them in a comprehensible way that seeks to put them into a theoretical, methodological and practical perspective.

The study has been conducted on various items of analysis through the integration of econometric modelling, forecasting, clustering analysis, and explainable machine learning. The final chapter sees the overall analysis of the importance of these strategy approaches in augmenting the knowledge about structural motivators and the future directions of renewable energy implementation. By so doing, it also tends to take cognisance of the limitations of big data comparative research and takes a critical look at the assumptions and trade-offs of such.

The chapter is outlined in the following way: Section 6.1 gives an overview of the most significant results of the research; further sections address the input to the knowledge and practice, the implications of the research outcomes on the policies, limitations of the research, and the opportunities for future research. Collectively, these parts give a holistic and inclusive conclusion to the thesis.

6.1 Summary of Key Findings

This thesis proposed formulating a data-driven and reproducible framework for analysing the uptake of renewable energy globally, and particularly solar photovoltaic and wind energy technologies. Combining an econometric analysis, a time-series forecasting, unsupervised machine learning, and explainable artificial intelligence, the study has offered a multi-layered insight into the structural drivers as well as predictive dynamics in the shaping of renewable electricity capacity in countries between 2000 and 2023.

The descriptive and exploratory analysis showed that the rate of growth in the adoption of renewable energy in the world has been faster over the last 20 years, especially since 2010. This has been accelerated by steep falls in the cost of technology, a rise in policy engagements in the wake of international climate agreements and the maturity of renewable energy markets. Nevertheless, this expansion has been extremely disparate, both inter-regional and inter-country. A few economies, such as China, the United States, Germany, and India, hold a comparatively high percentage of the world's installed capacity, with numerous economies of low-income and small sizes being at a very early adoption phase. Such skewed distribution carries the weight of the necessity to investigate heterogeneity in addition to examining global aggregates.

The econometric test offered some systematic data about the relationships of the renewable energy implementation. The overall result of the analyses was that GDP per capita has become a statistically significant predictor of renewable electricity capacity across all model specifications, supporting the importance of economic capacity, institutional strength and access to capital in facilitating large-scale renewable investment. There was also a positive correlation between population size, indicating that market size and aggregate energy demand provide an environment that would support the growth of renewables. The access rates of electricity were also positively correlated with the renewable capacity, reflecting the inherent importance of grid infrastructure and electrification in the process of integrating renewables.

On the other hand, the use of energy per capita had negative coefficients in multiple specifications. This observation indicates that the transition to a renewable energy system and fossil fuel-rich countries can encounter structural inertia due to their fossil fuel-based systems, the composition of industry, or consumption patterns. Of significance is that such findings do not necessarily mean that they are causal in a literal sense, but tend to draw consistent relationships that are consistent with the current theoretical and empirical view.

Fixed effects panel regression was the main econometric model that was used with the support of model diagnostics and specification tests. This option can be explained by the need to balance the impacts of unobserved and time-invariant country factors, including geography, history, orientation of policy and institutional quality. The fixed effects are more appropriate as this enhances the validity of relationships anticipated, as they reduce the omitted bias due to the influence of the same country-specific factors.

The predictive element of the thesis proved the existence of a significant difference in predictive performance in countries and methods of modelling. The classical time-series models, like the ARIMA, proved to be successful in comparatively more stable and mature energy systems, whereby the historical trends exhibit continuity and few structural breaks. Conversely, predicting renewable adoption in highly dynamic settings, including in emerging new economies experiencing a policy transformation and rapid expansions of infrastructure, became harder for all the models. Machine learning methods, especially XGBoost, were demonstrated to be possibly helpful at identifying the non-linear patterns, but were not always better than the more conventional methods, especially when learning on small amounts of time-series data.

Thanks to the clustering analysis, there were several typologies of countries based on the renewable capacity, the economic indicators, and the nature of the energy systems. The clusters coincide with the early and late adopters and leaders in the conceptual stages of an innovation diffusion process. This typological view that supplements regression-based findings emphasises similarities and differences that are not readily detected through marginal effects only.

Lastly, the analysis of explainability based on SHAP values increased the machine learning forecasts' transparency. The fact that most of the economically significant variables, namely GDP per capita, population, and electricity access, are utilised in the process of creating predictions draws confidence that the models have substantive relationships and are not spuriously correlated. It is especially significant in the case of policy-relevant applications where the trust and interpretation should be critical.

In summary, all five research questions have been addressed: RQ1 through descriptive trend analysis, RQ2 via panel regression, RQ3 through comparative forecasting, RQ4 using unsupervised clustering, and RQ5 by synthesizing policy-relevant insights from all prior results.

Research Question	Addressed By	Key Outcome
RQ1	Descriptive analysis (Sec 4.2)	Nonlinear global growth, concentrated in top adopters
RQ2	Fixed-effects regression (Sec 4.3)	GDP, population, and grid access are key drivers
RQ3	Forecasting comparison (Sec 4.4)	Model performance is context-dependent
RQ4	Clustering & PCA (Sec 5.4)	Adoption typologies reflect innovation diffusion stages
RQ5	Policy synthesis (Sec 6.3)	Tailored interventions needed by country archetype

Table 11: Research Objective / Question Mapping Table

6.2 Contributions to Knowledge and Practice

This thesis has a number of significant contributions in the theoretical, methodological and practical levels. Theoretically, the research would add to the body of literature on the adoption of renewable energy by offering recent, macro empirical data on more than twenty years and two hundred countries. Although most of the previous studies were done on a limited area or over a limited period, this thesis provides a global and longitudinal view, which represents long-run structural drivers and the rapid phases of acceleration. The results support previous theories of innovation diffusion, economic capacity and infrastructure preparedness, but also with existence of inequalities in the energy transition across the globe.

The methodological thesis contributes to the field of energy analytics by showing the benefits of combining both econometric and machine learning methods as one unified and consistent analytical chain. The study demonstrates that causal inference and prediction can actually be utilised in a complementary way as opposed to viewing them as opposite paradigms. Machine learning models are more flexible in prediction, and econometric models are more relevant in policy interpretation and elucidation. The clear distinction between explanatory and predictive aims enhances the clarity of analysis and eliminates the chance of mistakenly interpreting predictive success as evidence of causality.

The introduction of explainable artificial intelligence methods can be seen as another contribution to methodology. Using SHAP values on machine learning predictions, the thesis offers solutions to the typical criticisms of black-box models and shows that it is not necessary to compromise predictive accuracy in areas of education to keep the model transparent. This practice is especially applicable to applied research in energy and sustainability, where high stakes could be at stake based on model output.

Regarding practical and policy-related considerations, the findings provide some pieces of information. The linkage between renewable capacity and economic development is very strong, and this links to the necessity of special financial instruments like concessional finance, risk guarantees and international climate funds that should be made to enhance the use of renewables in the lower-income nations. The outcome of the clustering would indicate that countries are constrained and have various opportunities in different ways, meaning that homogeneous policy prescriptions cannot work. Rather, country-specific typology differentiated strategies can provide a more successful yield.

To practitioners and analysts, the reproducible computer framework that was created in this thesis could serve as a template for future research. Replicability is increased through the use of open-source tools, transparent data processing procedures and well-defined modelling pipelines, and extended by other researchers.

6.3 Policy Implications

There are many policy implications involving the empirical consequences of this thesis. To begin with, the role of economic capacity can be uniformly demonstrated by the need to balance renewable energy policies with the more general development policies. Education/ institutional quality/ financial market development investments indirectly contribute to renewable adoption through enhancing the mobilisation and management of capital-intensive infrastructure projects in a country.

Second, the fact that electricity access is positively related to renewable capacity indicates that the process of electrification and renewable implementation is something that should be done in a coherent fashion. Decentralised, off-grid renewable energy can be an interim step to wider grid integration in areas where the access rate is low enough to allow both energy access and decarbonisation to be advanced at the same time.

Third, some of the models positively relate renewable capacity to negative levels of energy use per capita; this is indicative of the difficulty of adapting high-use, fossil fuel-based systems. In the case of such nations, demand-side efficiency, industrial decarbonisation, and fossil fuel subsidy reform policy weapons could be valued just as highly as renewable production incentives.

The practical relevance of the forecasting results is also provided. The result that there is no one-size-fits-all model is a warning that no one technique of forecasting can benefit us. Ensemble techniques and scenario-based predictions, which factor in uncertainty and structural change, may be useful to policymakers and planners as opposed to point predictions.

Lastly, the clustering analysis provides a platform upon which peer learning and international collaboration can be accomplished. The member countries in the same cluster might have similar challenges and could be useful in sharing their best practices, policy making and technological solutions.

6.4 Limitations

In spite of its contributions, there are a number of limitations that can be identified within this study. First, the analysis is based on aggregate data at the country level that might conceal significant subnational processes. The application of renewable energy usually tends to be influenced by local resource endowments, local policies, as well as grid constraints, which are not capturable at the national level. Consequently, the results might be insufficient in their representation of national differences and local success stories.

Second, the policy variables have to be simplified in the representation. Even though indicators of a policy presence and persistence are used in the analysis, it fails to measure the variance in the quality of policy design, the effectiveness of the enforcement, and political stability. Such qualitative factors can be of vital importance in defining the real investment performance.

Third, global comparative research continues to have issues with the availability and quality of data. There were missing observations and inconsistencies in the measurement, and the strategies to clean and impute the data were to be conservative, and thus, they could inject uncertainty. Despite the strength tests performed, there might still be an effect of data limitations, especially on low-income and small nations.

Fourth, the analysis of the forecasting is limited to the comparatively short time series of the individual countries. In particular, machine learning models generally require large-scale datasets and cannot work effectively in thin or unsteady situations. Predicting based on structural breaks caused by abrupt policy changes or exogenous shocks is even more difficult.

Lastly, although the econometric models can regulate the unobserved time-invariant heterogeneity, they fail to provide strict experimental causality. The findings must be construed as either finding strong associations but not non-conclusive causal relationships.

The combination of restrictions explains why such empirical studies of renewable energy transitions on a large scale and cross-country level are inherently characterised by trade-offs. Although the application of aggregated data and simplified policy indicators allows high levels of broad comparability and global coverage, it cannot limit the capacity to record the localised dynamics, institutional peculiarities, and temporary disturbances. Lack of data and inconsistencies in measurements also indicate the difficulty in coming up with a full, comprehensive global analysis, especially in developing and small economies. Secondly, the limitations of predictions that come with short time series and structural breaks highlight the importance of being cautious about forecasting results. The greatest, but not the least, the econometric framework, which is used in this paper, has a strong nature as it manages the issue of unobserved heterogeneity; however, it still does not allow conclusive causal inference. Consequently, the results can only be considered as patterns and not as causal influences. By acknowledging these limitations, the importance of the study is not lost; rather, it will help offer a necessary background to explain the findings and offer a clear framework against which the results can be applied. Through the clear recognition of these limitations, the research has laid a clear basis that future studies may be developed based on the same, refined, and expanded the analysis approach based on richer data, better methodology and more granular points of view.

6.5 Future Research Recommendations

The research constraints of the study indicate that there are a number of feasible areas of future research. The integration of subnational and spatially explicit data is one of them. Local policy heterogeneity, resource endowment, and grid limitations can be better applied to the local level and capture a more intensive picture of adoption processes in the area.

It is also possible that future research incorporates more nuanced policy datasets that reflect the stringency of policies, financial size and quality of implementation. Quantitative indicators matched with qualitative analysis of policy can perhaps provide more insight into why comparable policies yield different results in different situations.

In terms of the methodology, the modelling framework can be extended methodologically to investigate heterogeneous treatment reactions of policies with causal machine learning methods, like causal forests or double machine learning. Such strategies would assist in determining the kind of nations that are most advantageous to the particular interventions.

On the forecasting side, future studies may seek to pursue ensemble models and scenario-based models that explicitly capture uncertainty, policy changes, and technological innovations. Additional predictive performance improvement could be possible by including exogenous variables like the investment flows, commodity prices or commitment to the climate policy.

Lastly, the analysis should be extended to other renewable methods, including hydro power, biomass, or new storage options, which would give a better and clearer view of the energy transition. With an improved availability of data, there might be additional opportunities for dynamic monitoring and evaluation using real-time/high-frequency data as well.

In general, such research directions are seen as prospects that indicate the dynamic and multidisciplinary character of renewable energy analytics. With the ongoing growth in the availability of data, computational capacity, and the complexity of policy, future research will be in a much better position to focus on the subtle, situational dynamics of renewable energy adoption. Combining spatially grained information, richer policy measures and superior causal modelling tools are especially promising in enhancing analytical accuracy and policy information. Simultaneously, more methodological development in the field of forecasting, including taking into account uncertainty, scenario analysis, and external economic factors, can lead to the increased reliability of forecasting in the energy systems that behave rapidly. The extension of the technological range of renewable sources and energy storage will also promote the comprehensive perspective of energy transitions. Collectively, such extensions would allow more dynamic, adaptive and evidence-based evaluation systems to assist all policy makers and stakeholders to develop specific interventions and track developments toward long-term sustainability objectives. In this regard, the current research can be seen as not the final step, but rather a base on which future studies can jump to a higher level of understanding of global renewable energy transitions.

6.6 Concluding Remarks

To summarise, this thesis is rigorous in terms of methods, and a thorough geographically wide examination of the adoption of renewable energy sources based on the model of data analysis and replicability. Through economic analysis, prediction, clustering, and explainable machine learning, the research provides both complementary information on what is driving the adoption of renewable energy sources and what is likely to happen in the near future.

The results highlight the point that despite the enormous movement in the world, the process of energy shifting is still uneven and predetermined by the structural roots of economies, infrastructures, and institutions. To manage these inequalities, a concerted policy effort, specific funding, and further methodological development of energy analytics will be needed.

In addition to the empirical results, this thesis indicates the increased role of interdisciplinary solutions that will close gaps between economics, data science, and energy policy. The complexifying and increasingly data-intensive nature of renewable energy systems in the future seems to be the driving force behind the change in the future decision-making process, which not only requires a further development of technologies but also a set of transparent, explainable, and reproducible analytical tools. This study is important to evolve the evidence-based energy policy analysis by showing how various quantitative approaches can be incorporated into a consistent computational process.

Finally, the shift to sustainable and low-carbon energy is not only a technical challenge but a social one as well. The analytical text and conclusions offered in this thesis will facilitate this change by offering policymakers, scholars, and professionals with sound knowledge and flexible instruments to comprehend, forecast, and inform the adoption of renewable energy in the world.

References

Abdelkader, S., Amissah, J., & Abdel-Rahim, O. (2024). Virtual power plants: an in-depth analysis of their advancements and importance as crucial players in modern power systems. *Energy, Sustainability and Society*, *14*(1), 52. https://link.springer.com/content/pdf/10.1186/s13705-024-00483-y.pdf

Abdullah, H., Sahudin, Z., & Bahrudin, N. Z. (2022). Short guides to the static panel data regression model estimator. *Asian Journal of Accounting and Finance*, *4*(4), 1-6.https://scholar.archive.org/work/mj3jexc2ofc6dgmut2oufhtt6m/access/wayback/https://myjms.mohe.gov.my/index.php/ajafin/article/download/20300/11089

Alhosani, A. (2023). *Development and Application of Forecasting Machine Learning Approach for Renewable Energy-Related Products' Prices* (Doctoral dissertation, Khalifa University of Science). https://khazna.ku.ac.ae/ws/portalfiles/portal/19572001/file

Alotaiq, A. (2024). Global low carbon transitions in the power sector: A machine learning clustering approach using archetypes. *Journal of Economy and Technology*, *2*, 95-127. https://www.sciencedirect.com/science/article/pii/S2949948824000155

Azizah, D. E. T. N., & Amalia, A. (2025). Legal Certainty and Investor Confidence: an Analysis of Indonesia's Capital Market Regulations. *International Journal of Law Dynamics Review*, *3*(1), 1-10. https://journal.nursyamcentre.com/index.php/IJLDR/article/download/61/60

Benti, N. E., Chaka, M. D., & Semie, A. G. (2023). Forecasting renewable energy generation with machine learning and deep learning: Current advances and prospects. *Sustainability*, *15*(9), 7087. https://www.mdpi.com/2071-1050/15/9/7087

Brecha, R. J., Ganti, G., Lamboll, R. D., Nicholls, Z., Hare, B., Lewis, J., ... & Gidden, M. J. (2022). Institutional decarbonization scenarios evaluated against the Paris Agreement 1.5 °C goal. *Nature communications*, *13*(1), 4304.https://www.nature.com/articles/s41467-022-31734-1.pdf

Brinkman, G., Bain, D., Buster, G., Draxl, C., Das, P., Ho, J., ... & Zhang, J. (2021). *The North American renewable integration study (NARIS): a US perspective* (No. NREL/TP-6A20-79224). National Renewable Energy Lab.(NREL), Golden, CO (United States). https://www.osti.gov/servlets/purl/1804701

Carpignano, A., Gerboni, R., Mezza, A., Pirri, C. F., Sacco, A., Sassone, D., ... E Viberti, D. (2023). Italian offshore platform and depleted reservoir conversion from the energy transition perspective. *Journal of Marine Science and Engineering*, *11*(8), 1544. https://www.mdpi.com/2077-1312/11/8/1544

Cline, C. (2025). Evaluating the Effect of Maine's 2019 Renewable Energy Legislation on Utility-Scale Solar. https://edepot.wur.nl/692897

Coronas, S., de la Hoz, J., Alonso, À., & Martín, H. (2022). 23 Years of development of the solar power generation sector in Spain: a comprehensive review of the period 1998–2020 from a regulatory perspective. *Energies*, *15*(4), 1593. https://www.mdpi.com/1996-1073/15/4/1593

Crijns-Graus, W., Wild, P., Amineh, M. P., Hu, J., & Yue, H. (2022). International comparison of research and investments in new renewable electricity technologies: A focus on the European Union and China. *Energies*, *15*(17), 6383.https://www.mdpi.com/1996-1073/15/17/6383

Elsaraiti, M., & Merabet, A. (2021). A comparative analysis of the ARIMA and LSTM predictive models and their effectiveness for predicting wind speed. *Energies*, *14*(20), 6782. https://www.mdpi.com/1996-1073/14/20/6782

Falcone, P. M. (2023). Sustainable energy policies in developing countries: a review of challenges and opportunities. *Energies*, *16*(18), 6682. https://www.mdpi.com/1996-1073/16/18/6682

Farhan, S. M., Yin, J., Chen, Z., & Memon, M. S. (2024). A comprehensive review of LiDAR applications in crop management for precision agriculture. *Sensors (Basel, Switzerland)*, *24*(16), 5409. https://pmc.ncbi.nlm.nih.gov/articles/PMC11360157/

Franco, G. C. (2025). *Seasonal electricity storage for renewable energy integration: technologies, challenges, and the case of Portugal* (Doctoral dissertation). https://repositorio.ulisboa.pt/bitstream/10400.5/101068/1/TM_Guilherme_Franco.pdf

Gajdzik, B., Wolniak, R., Nagaj, R., Grebski, W. W., & Romanyshyn, T. (2023). Barriers to renewable energy source (RES) installations as determinants of energy consumption in EU countries. *Energies*, *16*(21), 7364. https://www.mdpi.com/1996-1073/16/21/7364

Gargano, G. (2021). The bottom-up development model is a governance instrument for rural areas. The cases of four local action groups (LAGs) in the United Kingdom and Italy. *Sustainability*, *13*(16), 9123. https://www.mdpi.com/2071-1050/13/16/9123

Garrido-Herrero, M., Jaramillo-Moran, M. A., Carmona-Fernandez, D., & Ozcariz-Arraiza, I. M. (2024). The impact of photovoltaic self-consumption on the daily electricity demand in Spain: Definition of a model to estimate it. *Heliyon*, *10*(11). https://www.cell.com/heliyon/pdf/S2405-8440(24)08612-2.pdf

Ghasemi, A., Hosseinpanah, F., Kashfi, K., & Bahadoran, Z. (2025). Research hypothesis: a brief history, central role in scientific inquiry, and characteristics. *Addiction & Health*, *17*(1), 1-13.https://ahj.kmu.ac.ir/article_92855_a951a1d0100c47912145faf52da9590d.pdf

Gronow, A., Brockhaus, M., Di Gregorio, M., Karimo, A., & Ylä-Anttila, T. (2021). Policy learning as complex contagion: how social networks shape organizational beliefs in forest-based climate change mitigation. *Policy Sciences*, *54*(3), 529-556. https://link.springer.com/content/pdf/10.1007/s11077-021-09418-2.pdf

Gundersen, O. E. (2021). The fundamental principles of reproducibility. *Philosophical Transactions of the*

Royal Society A, *379*(2197), 20200210.https://royalsocietypublishing.org/doi/pdf/10.1098/rsta.2020.0210

Kacena, M. A., Plotkin, L. I., & Fehrenbacher, J. C. (2024). The use of artificial intelligence in writing scientific review articles. *Current Osteoporosis Reports*, *22*(1), 115-121. https://link.springer.com/content/pdf/10.1007/s11914-023-00852-0.pdf

Kalvenes, M. F., & Toftegaard, J. H. J. (2023). *Bidding for the Future: Efficient Auction Design during Financial Headwinds: A Comparative Study of UK and German Offshore Wind Auctions and Lessons for Norwegian Policymakers* (Master's thesis). https://openaccess.nhh.no/nhh-xmlui/bitstream/handle/11250/3131546/masterthesis.pdf?sequence=1

Khan, S. U. D., Wazeer, I., & Almutairi, Z. (2023). Comparative analysis of SAM and RETScreen tools for the case study of a 600 kW solar PV system installation in Riyadh, Saudi Arabia. *Sustainability*, *15*(6), 5381. https://www.mdpi.com/2071-1050/15/6/5381

Kontopoulou, V. I., Panagopoulos, A. D., Kakkos, I., & Matsopoulos, G. K. (2023). A review of ARIMA vs. machine learning approaches for time series forecasting in data-driven networks. *Future Internet*, *15*(8), 255. https://www.mdpi.com/1999-5903/15/8/255

Koutoudjian, G., Diniz, L., Cespedes, R., & UNDP, V. G. (2021). About IRENA The International Renewable Energy Agency (IRENA) is an intergovernmental organisation that supports countries in their transition to a sustainable energy future, and serves as the principal platform for international co-operation, a centre of excellence, and a repository of policy, technology, resource, and financial knowledge on renewable energy. *IRENA promotes the. IRENA promotes the*, *24*.https://energycentral.com/system/files/ece/nodes/504304/irena_rra_paraguay_2021.pdf

Kusuma, H. O. (2025). Use Of XGBoost Method For Very Short-Term Radiation Forecasting on Adaptive Solar Cells. *Journal of Intelligent Systems and Telecommunication*, *1*(2), 211-224. https://journal.unesa.ac.id/index.php/jistel/article/download/40698/13811

Lim, W. M. (2023). Philosophy of science and research paradigm for business research in the transformative age of automation, digitalization, hyperconnectivity, obligations, globalization, and sustainability. *Journal of Trade Science*, *11*(2-3), 3-30.https://www.emerald.com/jts/article-pdf/11/2-3/3/1719327/jts-07-2023-0015.pdf

Lu, Y., Karunasena, G., & Liu, C. (2022). A systematic literature review of non-compliance with low-carbon building regulations. *Energies*, *15*(24), 9266. https://www.mdpi.com/1996-1073/15/24/9266

Makowski, P. T., & Kajikawa, Y. (2021). Automation-driven innovation management? Toward the innovation-automation-strategy cycle. *Technological Forecasting and Social Change*, *168*, 120723. https://arxiv.org/pdf/2103.02395

Matushkin, D., Zaporozhets, A., Babak, V., Kulyk, M., & Denysov, V. (2025). Hourly Photovoltaic Power Generation Forecasting Using Exponential Smoothing: A Comparative Study Based on Operational Data.

https://www.preprints.org/frontend/manuscript/6f589c68d448f5e439cdcd17877fa3cc/download_pub

Matuszewska-Janica, A., Żebrowska-Suchodolska, D., Ala-Karvia, U., & Hozer-Koćmiel, M. (2021). Changes in electricity production from renewable energy sources in the European Union countries in 2005–2019. *Energies*, *14*(19), 6276. https://www.mdpi.com/1996-1073/14/19/6276

Mohamed, N. M. A., Kamal, K. M. M., & Binsuwadan, J. (2024). The adoption of renewable energy technologies by oil-producing countries: an inevitable outcome at a time of global challenges and demand for sustainable development. *Sustainability*, *16*(8), 3155. https://www.mdpi.com/2071-1050/16/8/3155

Mohammadi, A., Nápoles, G., & Salgueiro, Y. (2025). Deep Learning for Time-Series Forecasting with Exogenous Variables in Energy Consumption: A Performance and Interpretability Analysis. *IEEE Access*. https://ieeexplore.ieee.org/iel8/6287639/6514899/11005580.pdf

Osamika, D., Adelusi, B. S., Kelvin-Agwu, M. T. C., Mustapha, A. Y., Forkuo, A. Y., & Ikhalea, N. (2025). A Review of Data Visualization Tools and Techniques in Public Health: Enhancing Decision-Making through Analytics.https://worldscientificnews.com/wp-content/uploads/2025/05/WSN-203-2025-400-438.pdf

Rechsteiner, R. (2021). German energy transition (Energiewende) and what politicians can learn for environmental and climate policy. *Clean technologies and environmental policy*, *23*(2), 305-342. https://link.springer.com/content/pdf/10.1007/s10098-020-01939-3.pdf

Ridett, E., & Anderson, B. (2024). An assessment of the wider Hampshire distribution network capacity and potential constraint points for renewable generation (v1. 0): Establishing robust evidence for a pan-Hampshire 2025-2050 energy strategy project report. https://eprints.soton.ac.uk/490389/1/UoS-HCC-2023-WP3-Network-Capacity-Renewables-Constraints-v1.0.pdf

Saeed, S., & Siraj, T. (2024). Global Renewable Energy Infrastructure: Pathways to Carbon Neutrality and Sustainability. *Solar Energy and Sustainable Development Journal*, *13*(2), 183-203. https://jsesd-ojs.csers.ly/ojs/index.php/jsesd/article/view/243

Scheepens, D. R., Schicker, I., Hlaváčková-Schindler, K., & Plant, C. (2023). Adapting a deep convolutional RNN model with an imbalanced regression loss for improved spatio-temporal forecasting of extreme wind speed events in the short to medium range. *Geoscientific Model Development*, *16*(1), 251-270. https://gmd.copernicus.org/articles/16/251/2023/gmd-16-251-2023.pdf

Shah, S. S. (2024). Recent Trends in Renewable Energy Adoption and Their Economic Implications. *Journal of Environmental Science*, *1*, 100004. https://premierscience.com/wp-content/uploads/2024/10/pjes-24-311.pdf

Soto Nishimura, A., & Czaika, M. (2024). Exploring migration determinants: a meta-analysis of migration drivers and estimates. *Journal of International Migration and Integration*, *25*(2), 621-

643.https://link.springer.com/content/pdf/10.1007/s12134-023-01091-z.pdf

Swadi, M., Kadhim, D. J., Salem, M., Tuaimah, F. M., Majeed, A. S., & Alrubaie, A. J. (2024). Investigating and predicting the role of photovoltaic, wind, and hydrogen energies in sustainable global energy evolution. *Global Energy Interconnection*, *7*(4), 429-445.https://www.sciencedirect.com/science/article/pii/S2096511724000690/pdf?md5=eff79ff67e3a3421e6aab2553ce9b73a&pid=1-s2.0-S2096511724000690-main.pdf

Tabassum, S., Rahman, T., Islam, A. U., Rahman, S., Dipta, D. R., Roy, S., ... & Hossain, E. (2021). Solar energy in the United States: Development, challenges, and prospects. *Energies*, *14*(23), 8142. https://www.mdpi.com/1996-1073/14/23/8142

Vo, N. (2024). Assessing the impact of feed-in tariffs and renewable portfolio standards on the development of solar photovoltaic in Vietnam–opportunities and challenges. https://aaltodoc.aalto.fi/bitstreams/c1f65347-9114-46f6-a243-6394099555de/download

Yang, P. (2022). Urban expansion of Energiewende in Germany: a systematic bibliometric analysis and literature study. *Energy, Sustainability and Society*, *12*(1), 52. https://link.springer.com/content/pdf/10.1186/s13705-022-00373-1.pdf

Zhang, F., Gallagher, K. S., Myslikova, Z., Narassimhan, E., Bhandary, R. R., & Huang, P. (2021). From fossil to low carbon: The evolution of global public energy innovation. *Wiley Interdisciplinary Reviews: Climate Change*, *12*(6), e734. https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/wcc.734

Zhao, E., Sun, S., & Wang, S. (2022). New developments in wind energy forecasting with artificial intelligence and big data: A scientometric insight. *Data Science and Management*, *5*(2), 84-95.https://www.sciencedirect.com/science/article/pii/S2666764922000212

Zia, B., Rafiq, M., Saqib, S. E., & Atiq, M. (2022). Agricultural market competitiveness in the context of climate change: a systematic review. *Sustainability*, *14*(7), 3721. https://www.mdpi.com/2071-1050/14/7/3721

Salawu, R., Shamsuddin, A., Bolatitio, S., & Masibo, S. (2023). Theoretical and conceptual frameworks in research: Conceptual clarification. *European Chemical Bulletin*, *12*(12), 2103-2117.https://www.researchgate.net/profile/Aina-Obe-Shamsuddin/publication/374081258_THEORETICAL_AND_CONCEPTUAL_FRAMEWORKS_IN_RESEARCH_CONCEPTUAL_CLARIFICATION/links/650c631a82f01628f0361553/THEORETICAL-AND-CONCEPTUAL-FRAMEWORKS-IN-RESEARCH-CONCEPTUAL-CLARIFICATION.pdf

Leal, P. H., & Marques, A. C. (2022). The evolution of the environmental Kuznets curve hypothesis assessment: A literature review under a critical analysis perspective. *Heliyon*, *8*(11).https://www.cell.com/heliyon/pdf/S2405-8440(22)02809-2.pdf

Amin, R., Mathur, D., Ompong, D., & Zander, K. K. (2024). Integrating social aspects into energy system modelling through the lens of public perspectives: A review. *Energies*, *17*(23), 5880.https://www.mdpi.com/1996-1073/17/23/5880

Kyriazos, T., & Poga, M. (2024). Application of machine learning models in social sciences: managing nonlinear relationships. *Encyclopedia*, *4*(4), 1790-1805.https://www.mdpi.com/2673-8392/4/4/118

Strielkowski, W., Civín, L., Tarkhanova, E., Tvaronavičienė, M., & Petrenko, Y. (2021). Renewable energy in the sustainable development of electrical power sector: A review. *Energies*, *14*(24), 8240.https://www.mdpi.com/1996-1073/14/24/8240

Wang, L., He, Y., & Wu, R. (2024). The Green Engine of Growth: Assessing the Influence of Renewable Energy Consumption and Environmental Policy on China's Economic Sustainability. *Sustainability*, *16*(8), 3120.https://www.mdpi.com/2071-1050/16/8/3120

Tursunalieva, A., Alexander, D. L., Dunne, R., Li, J., Riera, L., & Zhao, Y. (2024). Making sense of machine learning: a review of interpretation techniques and their applications. *Applied Sciences*, *14*(2), 496.https://www.mdpi.com/2076-3417/14/2/496

de Mast, J., Steiner, S. H., Nuijten, W. P., & Kapitan, D. (2023). Analytical problem solving based on causal, correlational and deductive models. *The American Statistician*, *77*(1), 51-61.https://research.tue.nl/files/277763678/Analytical_Problem_Solving_Based_on_Causal_Correlational_and_Deductive_Models.pdf

Appendix

A1. STEP 1: Environment setup & data loading

```
import pandas as pd
import numpy as np


pd.set_option("display.max_columns", None)
pd.set_option("display.width", 120)

print("Libraries imported successfully.")
```

```
Libraries imported successfully.
```

```
IRENA_FILE = "IRENA_Electricity_Installed_Capacity_Solar_PV_Wind_2000_2023.csv"
WDI_FILE = "WorldBank_WDI_Core_Indicators_2000_2023.csv"
REN21_FILE = "REC_2021_Datapack.xlsx"
EI_FILE = "EI-Stats-Review-ALL-data.xlsx"

print("File paths defined.")
```

```
File paths defined.
```

```
irena_df = pd.read_csv(IRENA_FILE, encoding="latin1")
wdi_df = pd.read_csv(WDI_FILE, encoding="latin1")
ren21_df = pd.read_excel(REN21_FILE)
ei_df = pd.read_excel(EI_FILE)

print("Datasets loaded successfully.")
```

A2. Step 2: Dataset inspection & variable selection

```
irena_df.info()
wdi_df.info()
ren21_df.info()
ei_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
MultiIndex: 16129 entries, ('Country/area', 'Technology', 'Data Type', 'Grid connection', 'Year') to ('Zimbabwe', 'Offshore wind energy', 'Electri
city Installed Capacity (MW)', 'On-grid', '2023')
Data columns (total 1 columns):
 #   Column                                                                        Non-Null Count  Dtype
---  ------                                                                        --------------  -----
 0   Electricity statistics by Country/area, Technology, Data Type, Grid connection and Year  16129 non-null  object
dtypes: object(1)
memory usage: 231.6+ KB
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6384 entries, 0 to 6383
Data columns (total 8 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   country                 6384 non-null   object
 1   country_code            6360 non-null   object
 2   year                    6384 non-null   int64
 3   gdp_per_capita_usd      6194 non-null   float64
 4   population              6360 non-null   float64
 5   co2_emissions_kt        0 non-null      float64
 6   electricity_access_pct  6287 non-null   float64
 7   energy_use_per_capita   4656 non-null   float64
dtypes: float64(5), int64(1), object(2)
memory usage: 399.1+ KB
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 29 entries, 0 to 28
Data columns (total 9 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Unnamed: 0                            1 non-null      object
 1   Welcome to the REN21 REC 2021 Data Pack!  11 non-null      object
 2   Unnamed: 2                            8 non-null      object
 3   Unnamed: 3                            10 non-null     object
 4   Unnamed: 4                            8 non-null      object
 5   Unnamed: 5                            5 non-null      object
 6   Unnamed: 6                            8 non-null      object
 7   Unnamed: 7                            5 non-null      object
 8   Unnamed: 8                            4 non-null      object
dtypes: object(9)
memory usage: 2.2+ KB
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 121 entries, 0 to 120
Data columns (total 1 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Unnamed: 0  105 non-null    object
dtypes: object(1)
memory usage: 1.1+ KB
```

```
irena_df = pd.read_csv(
    "IRENA_Electricity_Installed_Capacity_Solar_PV_Wind_2000_2023.csv",
    encoding="latin1",
    index_col=[0, 1, 2, 3, 4]
)


irena_df.index.names
```

```
[6]: FrozenList([None, None, None, None, None])
```

```
[9]:

irena_df = pd.read_csv(
    "IRENA_Electricity_Installed_Capacity_Solar_PV_Wind_2000_2023.csv",
    encoding="latin1",
    index_col=[0, 1, 2, 3, 4]
)

irena_df.index = irena_df.index.set_names(
    ["country", "technology", "data_type", "grid_connection", "year"]
)

irena_df = irena_df.reset_index()
irena_df = irena_df.rename(columns={irena_df.columns[-1]: "capacity_mw"})

irena_df["year"] = pd.to_numeric(irena_df["year"], errors="coerce").astype("Int64")
irena_df["capacity_mw"] = pd.to_numeric(irena_df["capacity_mw"], errors="coerce")

irena_df = irena_df[
    irena_df["data_type"].str.contains("Installed Capacity", case=False, na=False) &
    irena_df["grid_connection"].str.contains("grid", case=False, na=False) &
    irena_df["technology"].str.contains(
        "solar|onshore wind|offshore wind",
        case=False,
        na=False
    )
]

irena_df = irena_df.pivot_table(
    index=["country", "year"],
    columns="technology",
    values="capacity_mw",
    aggfunc="sum"
).reset_index()

irena_df.columns = [
    "country",
    "year",
    "solar_pv_capacity",
    "onshore_wind_capacity",
    "offshore_wind_capacity"
]

irena_df.info()
display(irena_df.head())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5376 entries, 0 to 5375
Data columns (total 5 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   country                 5376 non-null   object
 1   year                    5376 non-null   Int64
 2   solar_pv_capacity       5376 non-null   float64
 3   onshore_wind_capacity   5376 non-null   float64
 4   offshore_wind_capacity  5376 non-null   float64
dtypes: Int64(1), float64(3), object(1)
memory usage: 215.4+ KB
```

	country	year	solar_pv_capacity	onshore_wind_capacity	offshore_wind_capacity
0	Afghanistan	2000	0.0	0.0	0.0
1	Afghanistan	2001	0.0	0.0	0.0
2	Afghanistan	2002	0.0	0.0	0.0
3	Afghanistan	2003	0.0	0.0	0.0
4	Afghanistan	2004	0.0	0.0	0.0

A3. STEP 3: Missingness Diagnostics & Country Filtering

```python
print("Rows:", irena_df.shape[0])
print("Countries:", irena_df["country"].nunique())
print("Years:", irena_df["year"].nunique())
print("Year range:", irena_df["year"].min(), "-", irena_df["year"].max())


missing_summary = irena_df.isna().mean().round(4) * 100
missing_summary


capacity_cols = [
    "solar_pv_capacity",
    "onshore_wind_capacity",
    "offshore_wind_capacity"
]

country_totals = irena_df.groupby("country")[capacity_cols].sum()
inactive_countries = country_totals[(country_totals.sum(axis=1) == 0)].index

len(inactive_countries)


irena_df_active = irena_df[~irena_df["country"].isin(inactive_countries)].copy()

print("Remaining countries:", irena_df_active["country"].nunique())
print("Remaining rows:", irena_df_active.shape[0])


data_quality = pd.DataFrame({
    "Metric": [
        "Initial observations",
        "Initial countries",
        "Final observations",
        "Final countries",
        "Missing values (%)",
        "Countries removed (zero capacity)"
    ],
    "Value": [
        irena_df.shape[0],
        irena_df["country"].nunique(),
        irena_df_active.shape[0],
        irena_df_active["country"].nunique(),
        missing_summary.mean(),
        len(inactive_countries)
    ]
})

data_quality


irena_df_active.info()
display(irena_df_active.head())
```

```
Rows: 5376
Countries: 224
Years: 24
Year range: 2000 - 2023
Remaining countries: 211
Remaining rows: 5064
<class 'pandas.core.frame.DataFrame'>
Index: 5064 entries, 0 to 5375
Data columns (total 5 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   country                 5064 non-null   object
 1   year                    5064 non-null   Int64
 2   solar_pv_capacity       5064 non-null   float64
 3   onshore_wind_capacity   5064 non-null   float64
 4   offshore_wind_capacity  5064 non-null   float64
dtypes: Int64(1), float64(3), object(1)
memory usage: 242.3+ KB
```

	country	year	solar_pv_capacity	onshore_wind_capacity	offshore_wind_capacity
0	Afghanistan	2000	0.0	0.0	0.0
1	Afghanistan	2001	0.0	0.0	0.0
2	Afghanistan	2002	0.0	0.0	0.0
3	Afghanistan	2003	0.0	0.0	0.0
4	Afghanistan	2004	0.0	0.0	0.0

A4. STEP 4: Integration with World Bank (WDI) Indicators

```python
wdi_df.info()
display(wdi_df.head())


wdi_df["year"] = pd.to_numeric(wdi_df["year"], errors="coerce").astype("Int64")
wdi_df["country"] = wdi_df["country"].astype(str)


wdi_df = wdi_df[
    (wdi_df["year"] >= irena_df_active["year"].min()) &
    (wdi_df["year"] <= irena_df_active["year"].max())
]


wdi_df = wdi_df[
    [
        "country",
        "year",
        "gdp_per_capita_usd",
        "population",
        "co2_emissions_kt",
        "electricity_access_pct",
        "energy_use_per_capita"
    ]
]


wdi_missing = wdi_df.isna().mean().round(4) * 100
wdi_missing


panel_df = pd.merge(
    irena_df_active,
    wdi_df,
    on=["country", "year"],
    how="left"
)


print("Final panel rows:", panel_df.shape[0])
print("Final panel countries:", panel_df["country"].nunique())

panel_df.info()
display(panel_df.head())


panel_df["total_wind_capacity"] = (
    panel_df["onshore_wind_capacity"] +
    panel_df["offshore_wind_capacity"]
)

panel_df["total_re_capacity"] = (
    panel_df["solar_pv_capacity"] +
    panel_df["total_wind_capacity"]
)

panel_df["log_gdp_per_capita"] = np.log(panel_df["gdp_per_capita_usd"])
panel_df["log_population"] = np.log(panel_df["population"])
panel_df["log_energy_use"] = np.log(panel_df["energy_use_per_capita"])


panel_df.isna().sum()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6384 entries, 0 to 6383
Data columns (total 8 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   country                 6384 non-null   object
 1   country_code            6360 non-null   object
 2   year                    6384 non-null   int64
 3   gdp_per_capita_usd      6194 non-null   float64
```

```
RangeIndex: 6384 entries, 0 to 6383
Data columns (total 8 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   country                6384 non-null   object 
 1   country_code           6360 non-null   object 
 2   year                   6384 non-null   int64  
 3   gdp_per_capita_usd     6194 non-null   float64
 4   population             6360 non-null   float64
 5   co2_emissions_kt       0 non-null      float64
 6   electricity_access_pct 6287 non-null   float64
 7   energy_use_per_capita  4656 non-null   float64
dtypes: float64(5), int64(1), object(2)
memory usage: 399.1+ KB
```

	country	country code	year	gdp per capita usd	population	co2 emissions kt	electricity access pct	energy use per capita
0	Afghanistan	AF	2000	174.930991	20130327.0	NaN	4.4	NaN
1	Afghanistan	AF	2001	138.706822	20284307.0	NaN	9.3	NaN
2	Afghanistan	AF	2002	178.954088	21378117.0	NaN	14.1	NaN
3	Afghanistan	AF	2003	198.871116	22733049.0	NaN	19.0	NaN
4	Afghanistan	AF	2004	221.763654	23560654.0	NaN	23.8	NaN

```
Final panel rows: 5064
Final panel countries: 211
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5064 entries, 0 to 5063
Data columns (total 10 columns):
 #   Column                  Non-Null Count  Dtype  
---  ------                  --------------  -----  
 0   country                 5064 non-null   object 
 1   year                    5064 non-null   Int64  
 2   solar_pv_capacity       5064 non-null   float64
 3   onshore_wind_capacity   5064 non-null   float64
 4   offshore_wind_capacity  5064 non-null   float64
 5   gdp_per_capita_usd      3704 non-null   float64
 6   population              3768 non-null   float64
 7   co2_emissions_kt        0 non-null      float64
 8   electricity_access_pct  3706 non-null   float64
 9   energy_use_per_capita   2780 non-null   float64
dtypes: Int64(1), float64(8), object(1)
memory usage: 400.7+ KB
```

	country	year	solar pv capacity	onshore wind capacity	offshore wind capacity	gdp per capita usd	population	co2 emissions kt	electricity access pct	e
0	Afghanistan	2000	0.0	0.0	0.0	174.930991	20130327.0	NaN	4.4	
1	Afghanistan	2001	0.0	0.0	0.0	138.706822	20284307.0	NaN	9.3	
2	Afghanistan	2002	0.0	0.0	0.0	178.954088	21378117.0	NaN	14.1	
3	Afghanistan	2003	0.0	0.0	0.0	198.871116	22733049.0	NaN	19.0	
4	Afghanistan	2004	0.0	0.0	0.0	221.763654	23560654.0	NaN	23.8	

11]:

```
country                      0
year                         0
solar_pv_capacity            0
onshore_wind_capacity        0
offshore_wind_capacity       0
gdp_per_capita_usd        1360
population                1296
co2_emissions_kt          5064
electricity_access_pct    1358
energy_use_per_capita     2284
total_wind_capacity          0
total_re_capacity            0
log_gdp_per_capita        1360
log_population            1296
log_energy_use            2284
dtype: int64
```

```
RangeIndex: 6384 entries, 0 to 6383
Data columns (total 8 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   country               6384 non-null   object
 1   country_code          6360 non-null   object
 2   year                  6384 non-null   int64
 3   gdp_per_capita_usd    6194 non-null   float64
 4   population            6360 non-null   float64
 5   co2_emissions_kt      0 non-null      float64
 6   electricity_access_pct  6287 non-null   float64
 7   energy_use_per_capita  4656 non-null   float64
dtypes: float64(5), int64(1), object(2)
memory usage: 399.1+ KB
```

	country	country code	year	gdp per capita usd	population	co2 emissions kt	electricity access pct	energy use per capita
0	Afghanistan	AF	2000	174.930991	20130327.0	NaN	4.4	NaN
1	Afghanistan	AF	2001	138.706822	20284307.0	NaN	9.3	NaN
2	Afghanistan	AF	2002	178.954088	21378117.0	NaN	14.1	NaN
3	Afghanistan	AF	2003	198.871116	22733049.0	NaN	19.0	NaN
4	Afghanistan	AF	2004	221.763654	23560654.0	NaN	23.8	NaN

```
Final panel rows: 5064
Final panel countries: 211
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5064 entries, 0 to 5063
Data columns (total 10 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   country                5064 non-null   object
 1   year                   5064 non-null   Int64
 2   solar_pv_capacity      5064 non-null   float64
 3   onshore_wind_capacity  5064 non-null   float64
 4   offshore_wind_capacity 5064 non-null   float64
 5   gdp_per_capita_usd     3704 non-null   float64
 6   population             3768 non-null   float64
 7   co2_emissions_kt       0 non-null      float64
 8   electricity_access_pct 3706 non-null   float64
 9   energy_use_per_capita  2780 non-null   float64
dtypes: Int64(1), float64(8), object(1)
memory usage: 400.7+ KB
```

	country	year	solar pv capacity	onshore wind capacity	offshore wind capacity	gdp per capita usd	population	co2 emissions kt	electricity access pct	e
0	Afghanistan	2000	0.0	0.0	0.0	174.930991	20130327.0	NaN	4.4	
1	Afghanistan	2001	0.0	0.0	0.0	138.706822	20284307.0	NaN	9.3	
2	Afghanistan	2002	0.0	0.0	0.0	178.954088	21378117.0	NaN	14.1	
3	Afghanistan	2003	0.0	0.0	0.0	198.871116	22733049.0	NaN	19.0	
4	Afghanistan	2004	0.0	0.0	0.0	221.763654	23560654.0	NaN	23.8	

```
country                      0
year                         0
solar_pv_capacity            0
onshore_wind_capacity        0
offshore_wind_capacity       0
gdp_per_capita_usd        1360
population                1296
co2_emissions_kt          5064
electricity_access_pct    1358
energy_use_per_capita     2284
total_wind_capacity          0
total_re_capacity            0
log_gdp_per_capita        1360
log_population            1296
log_energy_use            2284
dtype: int64
```

A5. STEP 5: Missing-Data Treatment & Imputation Rules

```python
panel_df = panel_df.drop(columns=["co2_emissions_kt"], errors="ignore")

panel_df = panel_df.sort_values(["country", "year"]).reset_index(drop=True)

panel_df["imp_gdp"] = panel_df["gdp_per_capita_usd"].isna().astype(int)
panel_df["imp_population"] = panel_df["population"].isna().astype(int)
panel_df["imp_electricity"] = panel_df["electricity_access_pct"].isna().astype(int)
panel_df["imp_energy_use"] = panel_df["energy_use_per_capita"].isna().astype(int)

panel_df["gdp_per_capita_usd"] = (
    panel_df
    .groupby("country")["gdp_per_capita_usd"]
    .transform(lambda x: x.interpolate(limit=2))
)

panel_df["population"] = (
    panel_df
    .groupby("country")["population"]
    .transform(lambda x: x.interpolate(limit=2))
)

panel_df["electricity_access_pct"] = (
    panel_df
    .groupby("country")["electricity_access_pct"]
    .transform(lambda x: x.interpolate(limit=2))
)

panel_df["energy_use_per_capita"] = (
    panel_df
    .groupby("country")["energy_use_per_capita"]
    .transform(lambda x: x.interpolate(limit=2))
)

panel_df["log_gdp_per_capita"] = np.log(panel_df["gdp_per_capita_usd"])
panel_df["log_population"] = np.log(panel_df["population"])
panel_df["log_energy_use"] = np.log(panel_df["energy_use_per_capita"])

imputation_summary = pd.DataFrame({
    "Variable": [
        "GDP per capita",
        "Population",
        "Electricity access",
        "Energy use per capita"
    ],
    "Imputed values": [
        panel_df["imp_gdp"].sum(),
        panel_df["imp_population"].sum(),
        panel_df["imp_electricity"].sum(),
        panel_df["imp_energy_use"].sum()
    ]
})

panel_df.isna().sum()
imputation_summary
panel_df.info()
panel_df.head()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5064 entries, 0 to 5063
Data columns (total 18 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   country                 5064 non-null   object
 1   year                    5064 non-null   Int64
 2   solar_pv_capacity       5064 non-null   float64
 3   onshore_wind_capacity   5064 non-null   float64
 4   offshore_wind_capacity  5064 non-null   float64
 5   gdp_per_capita_usd      3710 non-null   float64
 6   population              3768 non-null   float64
 7   electricity_access_pct  3706 non-null   float64
 8   energy_use_per_capita   2874 non-null   float64
 9   total_wind_capacity     5064 non-null   float64
 10  total_re_capacity       5064 non-null   float64
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5064 entries, 0 to 5063
Data columns (total 18 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   country                 5064 non-null   object
 1   year                    5064 non-null   Int64
 2   solar_pv_capacity       5064 non-null   float64
 3   onshore_wind_capacity   5064 non-null   float64
 4   offshore_wind_capacity  5064 non-null   float64
 5   gdp_per_capita_usd      3710 non-null   float64
 6   population              3768 non-null   float64
 7   electricity_access_pct  3706 non-null   float64
 8   energy_use_per_capita   2874 non-null   float64
 9   total_wind_capacity     5064 non-null   float64
 10  total_re_capacity       5064 non-null   float64
 11  log_gdp_per_capita      3710 non-null   float64
 12  log_population          3768 non-null   float64
 13  log_energy_use          2874 non-null   float64
 14  imp_gdp                 5064 non-null   int64
 15  imp_population          5064 non-null   int64
 16  imp_electricity         5064 non-null   int64
 17  imp_energy_use          5064 non-null   int64
dtypes: Int64(1), float64(12), int64(4), object(1)
memory usage: 717.2+ KB
```

	country	year	solar_pv_capacity	onshore_wind_capacity	offshore_wind_capacity	gdp_per_capita_usd	population	electricity_access_pct	energy_use_per_capit
0	Afghanistan	2000	0.0	0.0	0.0	174.930991	20130327.0	4.4	NaI
1	Afghanistan	2001	0.0	0.0	0.0	138.706822	20284307.0	9.3	NaI
2	Afghanistan	2002	0.0	0.0	0.0	178.954088	21378117.0	14.1	NaI
3	Afghanistan	2003	0.0	0.0	0.0	198.871116	22733049.0	19.0	NaI
4	Afghanistan	2004	0.0	0.0	0.0	221.763654	23560654.0	23.8	NaI

A6. STEP 6: Exploratory Data Analysis (EDA)

[25]:

```
desc_stats = panel_df[
    [
        "solar_pv_capacity",
        "onshore_wind_capacity",
        "offshore_wind_capacity",
        "total_re_capacity",
        "gdp_per_capita_usd",
        "population",
        "electricity_access_pct",
        "energy_use_per_capita"
    ]
].describe()

desc_stats
```

[25]:

	solar pv capacity	onshore wind capacity	offshore wind capacity	total re capacity	gdp per capita usd	population	electricity access pct	energy use per
count	5064.000000	5064.000000	5064.000000	5.064000e+03	3710.000000	3.768000e+03	3706.000000	2874.
mean	70.661291	1513.442289	1290.458118	2.874562e+03	13937.068720	3.670099e+07	81.140340	2378.
std	956.433856	12513.188270	13584.681056	2.626029e+04	19676.038081	1.498006e+08	29.110452	2964.
min	0.000000	0.000000	0.000000	0.000000e+00	109.593814	9.544000e+03	0.800000	9.
25%	0.000000	0.000000	0.000000	0.000000e+00	1580.563837	1.219356e+06	70.325000	560.
50%	0.000000	0.000000	0.220000	2.000000e+00	5141.779434	6.825680e+06	98.900000	1381.
75%	0.000000	41.900000	23.000000	1.178200e+02	18629.317824	2.073615e+07	100.000000	2856.
max	37290.000000	404050.000000	608920.000000	1.050260e+06	134965.815442	1.438070e+09	100.000000	21557.

[26]:

```
global_trends = (
    panel_df
    .groupby("year")[
        [
            "solar_pv_capacity",
            "onshore_wind_capacity",
            "offshore_wind_capacity",
            "total_re_capacity"
        ]
    ]
    .sum()
    .reset_index()
)

global_trends
```

[26]:

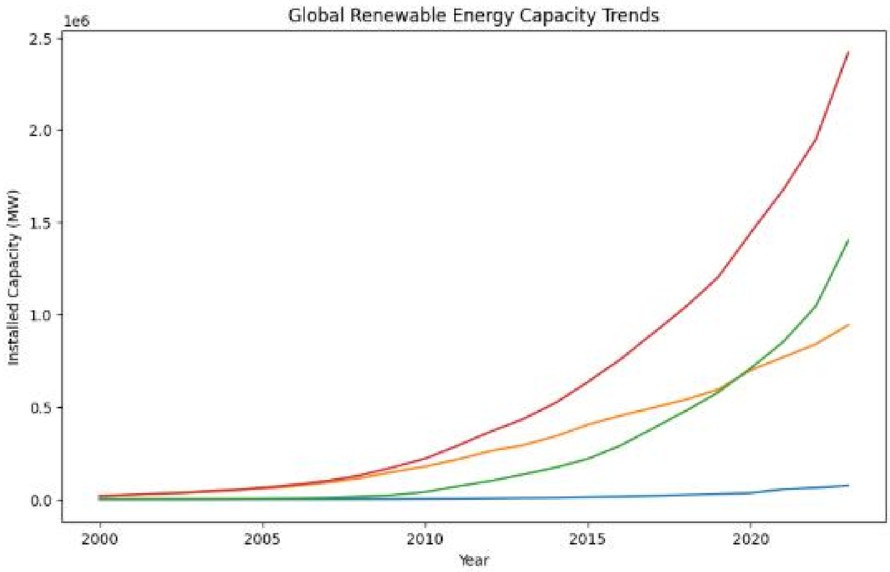
year	solar pv capacity	onshore wind capacity	offshore wind capacity	total re capacity

[26]:

	year	solar_pv_capacity	onshore_wind_capacity	offshore_wind_capacity	total_re_capacity
0	2000	66.95	16881.33	755.69	17703.97
1	2001	75.95	23872.38	1022.80	24971.13
2	2002	239.95	30474.36	1341.02	32055.33
3	2003	510.55	38142.98	1847.47	40501.00
4	2004	595.75	47057.51	2915.31	50568.57
5	2005	685.75	57770.35	4390.35	62846.45
6	2006	883.75	72252.81	5911.64	79048.20
7	2007	1094.25	90411.65	8265.92	99771.82
8	2008	1442.45	114019.74	14444.86	129907.05
9	2009	2133.55	147956.61	22483.40	172573.56
10	2010	3055.35	177711.64	39732.76	220499.75
11	2011	3786.61	216246.13	70189.34	290222.08
12	2012	5344.75	261361.27	99017.28	365723.30
13	2013	7182.05	292277.12	134040.27	433499.44
14	2014	8502.57	340639.79	172168.74	521311.10
15	2015	11732.87	404246.51	219262.06	635241.44
16	2016	14377.25	452487.45	290561.22	757425.92
17	2017	18804.90	495687.33	384045.12	898537.35
18	2018	23559.60	539444.58	477686.52	1040690.70
19	2019	28263.30	594236.85	578978.21	1201478.36
20	2020	34339.95	698422.70	709092.74	1441855.39
21	2021	54268.38	769389.53	851793.59	1675451.50
22	2022	62627.09	839699.44	1044311.41	1946637.94
23	2023	74255.21	943381.69	1400622.19	2418259.09

[27]:

```
import matplotlib.pyplot as plt

plt.figure(figsize=(10, 6))
plt.plot(global_trends["year"], global_trends["solar_pv_capacity"])
plt.plot(global_trends["year"], global_trends["onshore_wind_capacity"])
plt.plot(global_trends["year"], global_trends["offshore_wind_capacity"])
plt.plot(global_trends["year"], global_trends["total_re_capacity"])
plt.xlabel("Year")
plt.ylabel("Installed Capacity (MW)")
plt.title("Global Renewable Energy Capacity Trends")
plt.show()
```




```
top_countries = (
    panel_df
    .groupby("country")["total_re_capacity"]
    .max()
    .sort_values(ascending=False)
    .head(15)
)

top_countries
```

```
country
China                                                        1050260.00
United States of America (the)                                285966.90
Germany                                                       144368.00
India                                                         115498.56
Japan                                                          92437.04
Brazil                                                         67049.33
Spain                                                          60448.11
United Kingdom of Great Britain and Northern Ireland (the)     46438.21
Australia                                                      43579.94
Italy                                                          41658.70
France                                                         40157.34
Netherlands (Kingdom of the)                                   32008.76
Poland                                                         25770.85
Republic of Korea (the)                                        25692.94
```

```
Australia                                    43579.94
Italy                                        41658.70
France                                       40157.34
Netherlands (Kingdom of the)                 32088.76
Poland                                       25770.85
Republic of Korea (the)                      25692.94
Viet Nam                                     24464.42
Name: total_re_capacity, dtype: float64
```

9]:

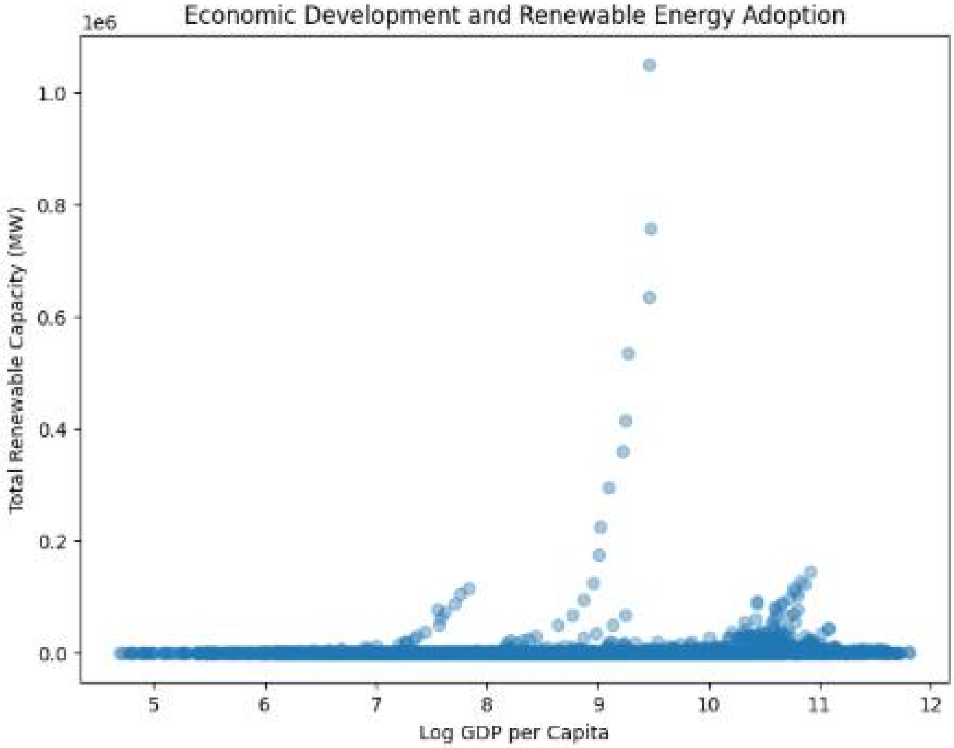
```
corr_matrix = panel_df[
    [
        "log_gdp_per_capita",
        "log_population",
        "log_energy_use",
        "electricity_access_pct",
        "total_re_capacity"
    ]
].corr()

corr_matrix
```

9]:

	log gdp per capita	log population	log energy use	electricity access pct	total re capacity
log gdp per capita	1.000000	-0.253789	0.845150	0.728215	0.080116
log population	-0.253789	1.000000	-0.198849	-0.165896	0.185957
log energy use	0.845150	-0.198849	1.000000	0.689415	0.068859
electricity access pct	0.728215	-0.165896	0.689415	1.000000	0.066821
total re capacity	0.080116	0.185957	0.068859	0.066821	1.000000

8]:

```
plt.figure(figsize=(8, 6))
plt.scatter(panel_df["log_gdp_per_capita"], panel_df["total_re_capacity"], alpha=0.4)
plt.xlabel("Log GDP per Capita")
plt.ylabel("Total Renewable Capacity (MW)")
plt.title("Economic Development and Renewable Energy Adoption")
plt.show()
```



[31]:

```
eda_summary = pd.DataFrame({
    "Metric": [
        "Total observations",
        "Countries",
        "Years covered",
        "Mean total RE capacity (MW)",
        "Median total RE capacity (MW)"
    ],
    "Value": [
        panel_df.shape[0],
        panel_df["country"].nunique(),
        panel_df["year"].nunique(),
        panel_df["total_re_capacity"].mean(),
        panel_df["total_re_capacity"].median()
    ]
})

eda_summary
```

[31]:

	Metric	Value
0	Total observations	5064.000000
1	Countries	211.000000
2	Years covered	24.000000
3	Mean total RE capacity (MW)	2874.561698
4	Median total RE capacity (MW)	2.000000

A7. Step 7: Econometric Modelling (Panel Regression)

```
!pip install linearmodels
```

```
Collecting linearmodels
  Downloading linearmodels-7.0-cp312-cp312-win_amd64.whl.metadata (10 kB)
Requirement already satisfied: numpy<3,>=1.22.3 in d:\python\lib\site-packages (from linearmodels) (2.1.3)
Requirement already satisfied: pandas>=1.4.0 in d:\python\lib\site-packages (from linearmodels) (2.3.1)
Requirement already satisfied: scipy>=1.8.0 in d:\python\lib\site-packages (from linearmodels) (1.13.1)
Requirement already satisfied: statsmodels>=0.13.0 in d:\python\lib\site-packages (from linearmodels) (0.14.4)
Requirement already satisfied: mypy_extensions>=0.4 in d:\python\lib\site-packages (from linearmodels) (1.0.0)
Collecting pyhdfe>=0.1 (from linearmodels)
  Downloading pyhdfe-0.2.0-py3-none-any.whl.metadata (4.0 kB)
Collecting formulaic>=1.2.1 (from linearmodels)
  Downloading formulaic-1.2.1-py3-none-any.whl.metadata (7.0 kB)
Collecting interface-meta>=1.2.0 (from formulaic>=1.2.1->linearmodels)
  Downloading interface_meta-1.3.0-py3-none-any.whl.metadata (6.7 kB)
Requirement already satisfied: narwhals>=1.17 in d:\python\lib\site-packages (from formulaic>=1.2.1->linearmodels) (1.30.0)
Requirement already satisfied: typing-extensions>=4.2.0 in d:\python\lib\site-packages (from formulaic>=1.2.1->linearmodels) (4.12.2)
Requirement already satisfied: wrapt>=1.0 in d:\python\lib\site-packages (from formulaic>=1.2.1->linearmodels) (1.17.2)
Requirement already satisfied: python-dateutil>=2.8.2 in d:\python\lib\site-packages (from pandas>=1.4.0->linearmodels) (2.9.0.post0)
Requirement already satisfied: pytz>=2020.1 in d:\python\lib\site-packages (from pandas>=1.4.0->linearmodels) (2025.1)
Requirement already satisfied: tzdata>=2022.7 in d:\python\lib\site-packages (from pandas>=1.4.0->linearmodels) (2025.1)
Requirement already satisfied: six>=1.5 in d:\python\lib\site-packages (from python-dateutil>=2.8.2->pandas>=1.4.0->linearmodels) (1.17.0)
Requirement already satisfied: patsy>=0.5.6 in d:\python\lib\site-packages (from statsmodels>=0.13.0->linearmodels) (1.0.1)
Requirement already satisfied: packaging>=21.3 in d:\python\lib\site-packages (from statsmodels>=0.13.0->linearmodels) (24.0)
Downloading linearmodels-7.0-cp312-cp312-win_amd64.whl (1.5 MB)
   ---------------------------------------- 0.0/1.5 MB ? eta -:--:--
   --------------------------- ------------ 1.0/1.5 MB 25.4 MB/s eta 0:00:01
   ---------------------------------------- 1.5/1.5 MB 3.5 MB/s eta 0:00:00
Downloading formulaic-1.2.1-py3-none-any.whl (117 kB)
Downloading interface_meta-1.3.0-py3-none-any.whl (14 kB)
Downloading pyhdfe-0.2.0-py3-none-any.whl (19 kB)
Installing collected packages: interface-meta, pyhdfe, formulaic, linearmodels

   -------------------- ------------------- 2/4 [formulaic]
   ------------------------------ --------- 3/4 [linearmodels]
   ------------------------------ --------- 3/4 [linearmodels]
   ------------------------------ --------- 3/4 [linearmodels]
   ------------------------------ --------- 3/4 [linearmodels]
   ---------------------------------------- 4/4 [linearmodels]

Successfully installed formulaic-1.2.1 interface-meta-1.3.0 linearmodels-7.0 pyhdfe-0.2.0

[notice] A new release of pip is available: 25.1.1 -> 25.3
[notice] To update, run: python.exe -m pip install --upgrade pip
```

```python
import numpy as np
import pandas as pd
import statsmodels.api as sm
from linearmodels.panel import PanelOLS, RandomEffects
from linearmodels.panel import compare
from statsmodels.stats.outliers_influence import variance_inflation_factor

panel_df["log_total_re_capacity"] = np.log(panel_df["total_re_capacity"] + 1)

if not isinstance(panel_df.index, pd.MultiIndex):
    panel_df = panel_df.set_index(["country", "year"])

X = panel_df[
    [
        "log_gdp_per_capita",
        "log_population",
        "electricity_access_pct",
        "log_energy_use"
    ]
]

X = sm.add_constant(X)

y = panel_df["log_total_re_capacity"]
```

```python
    "Fixed Effects": fe_model,
    "Random Effects": re_model
})

X_vif = X.replace([np.inf, -np.inf], np.nan).dropna()

vif_data = pd.DataFrame({
    "Variable": X_vif.columns,
    "VIF": [
        variance_inflation_factor(X_vif.values, i)
        for i in range(X_vif.shape[1])
    ]
})

vif_data


model_summary = pd.DataFrame({
    "Model": ["Pooled OLS", "Fixed Effects", "Random Effects"],
    "R_squared": [
        pooled_ols.rsquared,
        fe_model.rsquared,
        re_model.rsquared
    ]
})

pooled_ols.summary(), fe_model.summary, re_model.summary, comparison, vif_data, model_summary
```

```
D:\Python\Lib\site-packages\linearmodels\panel\model.py:1258: MissingValueWarning:
Inputs contain missing values. Dropping rows with missing observations.
  super().__init__(dependent, exog, weights=weights, check_rank=check_rank)
D:\Python\Lib\site-packages\linearmodels\panel\model.py:2751: MissingValueWarning:
Inputs contain missing values. Dropping rows with missing observations.
  super().__init__(dependent, exog, weights=weights, check_rank=check_rank)
```

```
(<class 'statsmodels.iolib.summary.Summary'>
 """
                              OLS Regression Results
 ==============================================================================
 Dep. Variable:     log_total_re_capacity   R-squared:                       0.595
 Model:                               OLS   Adj. R-squared:                  0.595
 Method:                    Least Squares   F-statistic:                     1387.
 Date:                   Wed, 17 Dec 2025   Prob (F-statistic):               0.00
 Time:                           23:53:45   Log-Likelihood:                -6225.9
 No. Observations:                   2850   AIC:                         1.246e+04
 Df Residuals:                       2845   BIC:                         1.249e+04
 Df Model:                              4
 Covariance Type:                     HC3
 =========================================================================================
                             coef    std err          z      P>|z|      [0.025      0.975]
 -----------------------------------------------------------------------------------------
 const                   -24.0885      0.490    -49.157      0.000     -25.049     -23.128
 log_gdp_per_capita        2.0757      0.048     43.093      0.000       1.981       2.170
 log_population            1.0328      0.025     41.512      0.000       0.984       1.082
 electricity_access_pct    0.0135      0.002      6.815      0.000       0.010       0.017
 log_energy_use           -1.1472      0.066    -17.466      0.000      -1.276      -1.018
 ==============================================================================
 Omnibus:                      106.431   Durbin-Watson:                   0.173
 Prob(Omnibus):                  0.000   Jarque-Bera (JB):               69.109
 Skew:                          -0.256   Prob(JB):                     9.84e-16
 Kurtosis:                       2.435   Cond. No.                     1.20e+03
 ==============================================================================

 Notes:
 [1] Standard Errors are heteroscedasticity robust (HC3)
 [2] The condition number is large, 1.2e+03. This might indicate that there are
 strong multicollinearity or other numerical problems.
 """,
 <class 'linearmodels.compat.statsmodels.Summary'>
 """
                           PanelOLS Estimation Summary
 ================================================================================
 Dep. Variable:     log_total_re_capacity   R-squared:                        0.4716
 Estimator:                      PanelOLS   R-squared (Between):             -2.6336
 No. Observations:                   2850   R-squared (Within):               0.4716
 Date:                   Wed, Dec 17 2025   R-squared (Overall):             -0.8898
```

```
==============================================================================
Dep. Variable:     log_total_re_capacity   R-squared:                       0.595
Model:                               OLS   Adj. R-squared:                  0.595
Method:                    Least Squares   F-statistic:                     1387.
Date:                   Wed, 17 Dec 2025   Prob (F-statistic):               0.00
Time:                           23:53:45   Log-Likelihood:                -6225.9
No. Observations:                   2850   AIC:                         1.246e+04
Df Residuals:                       2845   BIC:                         1.249e+04
Df Model:                              4
Covariance Type:                     HC3
=========================================================================================
                           coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------
const                  -24.0885      0.490    -49.157      0.000     -25.049     -23.128
log_gdp_per_capita       2.0757      0.048     43.093      0.000       1.981       2.170
log_population           1.0328      0.025     41.512      0.000       0.984       1.082
electricity_access_pct   0.0135      0.002      6.815      0.000       0.010       0.017
log_energy_use          -1.1472      0.066    -17.466      0.000      -1.276      -1.018
==============================================================================
Omnibus:                      106.431   Durbin-Watson:                   0.173
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               69.109
Skew:                          -0.256   Prob(JB):                     9.84e-16
Kurtosis:                       2.435   Cond. No.                     1.20e+03
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)
[2] The condition number is large, 1.2e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
""",
<class 'linearmodels.compat.statsmodels.Summary'>
"""
                          PanelOLS Estimation Summary
================================================================================
Dep. Variable:     log_total_re_capacity   R-squared:                        0.4716
Estimator:                      PanelOLS   R-squared (Between):             -2.6336
No. Observations:                   2850   R-squared (Within):               0.4716
Date:                   Wed, Dec 17 2025   R-squared (Overall):             -0.8098
Time:                           23:53:45   Log-likelihood                   -5153.3
Cov. Estimator:                Clustered
                                           F-statistic:                      605.12
Entities:                            134   P-value                           0.0000
Avg Obs:                          21.269   Distribution:                  F(4,2712)
Min Obs:                          6.0000
Max Obs:                          24.000   F-statistic (robust):             68.808
                                           P-value                           0.0000
Time periods:                         24   Distribution:                  F(4,2712)
Avg Obs:                          118.75
Min Obs:                          113.00
Max Obs:                          133.00

                                   Parameter Estimates
==========================================================================================
                        Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
------------------------------------------------------------------------------------------
const                     -67.883     14.216    -4.7752     0.0000     -95.758     -40.008
log_gdp_per_capita         2.3408     0.2824     8.2882     0.0000      1.7870      2.8946
log_population             3.4177     0.8824     3.8733     0.0001      1.6875      5.1479
electricity_access_pct     0.0233     0.0115     2.0191     0.0436      0.0007      0.0459
log_energy_use            -0.8259     0.7507    -1.1001     0.2714     -2.2980      0.6462
==========================================================================================

F-test for Poolability: 22.896
P-value: 0.0000
Distribution: F(133,2712)

Included effects: Entity
""",
<class 'linearmodels.compat.statsmodels.Summary'>
"""
                        RandomEffects Estimation Summary
================================================================================
Dep. Variable:     log_total_re_capacity   R-squared:                        0.4643
Estimator:                 RandomEffects   R-squared (Between):              0.6271
No. Observations:                   2850   R-squared (Within):               0.4558
Date:                   Wed, Dec 17 2025   R-squared (Overall):              0.5589
Time:                           23:53:45   Log-likelihood                   -5279.6
```

```
=================================================================================
Dep. Variable:      log_total_re_capacity   R-squared:                        0.4643
Estimator:                  RandomEffects   R-squared (Between):              0.6271
No. Observations:                    2850   R-squared (Within):               0.4558
Date:                    Wed, Dec 17 2025   R-squared (Overall):              0.5589
Time:                            23:53:45   Log-likelihood                   -5279.6
Cov. Estimator:                 Clustered
                                            F-statistic:                      616.54
Entities:                             134   P-value                           0.0000
Avg Obs:                           21.269   Distribution:                  F(4,2845)
Min Obs:                           6.0000
Max Obs:                           24.000   F-statistic (robust):             120.06
                                            P-value                           0.0000
Time periods:                          24   Distribution:                  F(4,2845)
Avg Obs:                           118.75
Min Obs:                           113.00
Max Obs:                           133.00

                             Parameter Estimates
=========================================================================================
                        Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
-----------------------------------------------------------------------------------------
const                     -27.892     1.8170    -15.351     0.0000     -31.455     -24.330
log_gdp_per_capita         2.5584     0.2155     11.871     0.0000      2.1358      2.9810
log_population             1.1869     0.0899     13.199     0.0000      1.0106      1.3632
electricity_access_pct     0.0298     0.0080     3.7110     0.0002      0.0141      0.0456
log_energy_use            -1.7251     0.2786    -6.1923     0.0000     -2.2713     -1.1788
=========================================================================================
""",
                         Model Comparison
============================================================================
                                     Fixed Effects          Random Effects
----------------------------------------------------------------------------
Dep. Variable                log_total_re_capacity   log_total_re_capacity
Estimator                                 PanelOLS           RandomEffects
No. Observations                              2850                    2850
Cov. Est.                                Clustered               Clustered
R-squared                                   0.4716                  0.4643
R-Squared (Within)                          0.4716                  0.4558
R-Squared (Between)                        -2.6336                  0.6271
R-Squared (Overall)                        -0.8098                  0.5589
F-statistic                                 685.12                  616.54
P-value (F-stat)                            0.0000                  0.0000
========================    ======================  ======================
const                                      -67.883                 -27.892
                                         (-4.7752)               (-15.351)
log_gdp_per_capita                          2.3408                  2.5584
                                          (8.2882)                (11.871)
log_population                              3.4177                  1.1869
                                          (3.8733)                (13.199)
electricity_access_pct                      0.0233                  0.0298
                                          (2.0191)                (3.7110)
log_energy_use                             -0.8259                 -1.7251
                                         (-1.1001)               (-6.1923)
========================== ======================= =======================
Effects                                     Entity
----------------------------------------------------------------------------

T-stats reported in parentheses
PanelModelComparison, id: 0x1f3486a3590,
                Variable         VIF
0                  const  176.675914
1     log_gdp_per_capita    3.952011
2         log_population    1.075614
3 electricity_access_pct    2.084202
4         log_energy_use    3.746862,
             Model  R_squared
0       Pooled OLS   0.595382
1    Fixed Effects   0.471600
2   Random Effects   0.464336)
```

A8. Step 8: Forecasting

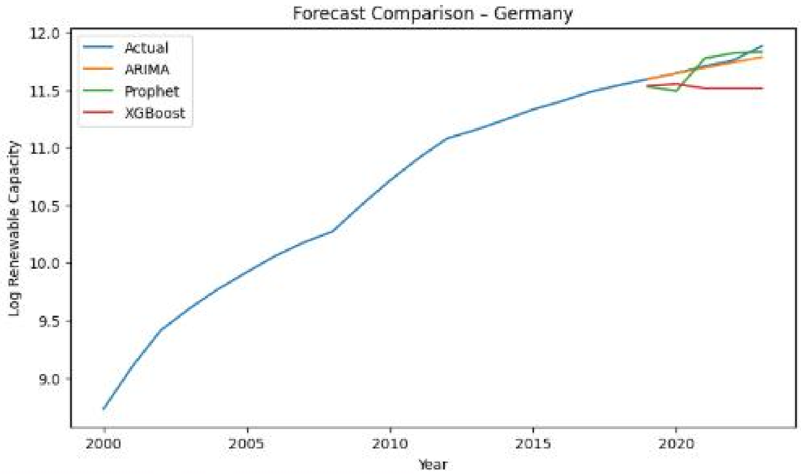
```
!pip install prophet
```

```
Collecting prophet
  Downloading prophet-1.2.1-py3-none-win_amd64.whl.metadata (3.6 kB)
Collecting cmdstanpy>=1.0.4 (from prophet)
  Downloading cmdstanpy-1.3.0-py3-none-any.whl.metadata (4.2 kB)
Requirement already satisfied: numpy>=1.15.4 in d:\python\lib\site-packages (from prophet) (2.1.3)
Requirement already satisfied: matplotlib>=2.0.0 in d:\python\lib\site-packages (from prophet) (3.10.5)
Requirement already satisfied: pandas>=1.0.4 in d:\python\lib\site-packages (from prophet) (2.3.1)
Collecting holidays<1,>=0.25 (from prophet)
  Downloading holidays-0.87-py3-none-any.whl.metadata (50 kB)
Requirement already satisfied: tqdm>=4.36.1 in d:\python\lib\site-packages (from prophet) (4.67.1)
Requirement already satisfied: importlib_resources in d:\python\lib\site-packages (from prophet) (6.5.2)
Requirement already satisfied: python-dateutil<3,>=2.9.0.post0 in d:\python\lib\site-packages (from holidays<1,>=0.25->prophet) (2.9.0.post0)
Requirement already satisfied: six>=1.5 in d:\python\lib\site-packages (from python-dateutil<3,>=2.9.0.post0->holidays<1,>=0.25->prophet) (1.17.0)
Collecting stanio<2.0.0,>=0.4.0 (from cmdstanpy>=1.0.4->prophet)
  Downloading stanio-0.5.1-py3-none-any.whl.metadata (1.6 kB)
Requirement already satisfied: contourpy>=1.0.1 in d:\python\lib\site-packages (from matplotlib>=2.0.0->prophet) (1.3.3)
Requirement already satisfied: cycler>=0.10 in d:\python\lib\site-packages (from matplotlib>=2.0.0->prophet) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in d:\python\lib\site-packages (from matplotlib>=2.0.0->prophet) (4.59.0)
Requirement already satisfied: kiwisolver>=1.3.1 in d:\python\lib\site-packages (from matplotlib>=2.0.0->prophet) (1.4.8)
Requirement already satisfied: packaging>=20.0 in d:\python\lib\site-packages (from matplotlib>=2.0.0->prophet) (24.0)
Requirement already satisfied: pillow>=8 in d:\python\lib\site-packages (from matplotlib>=2.0.0->prophet) (11.3.0)
Requirement already satisfied: pyparsing>=2.3.1 in d:\python\lib\site-packages (from matplotlib>=2.0.0->prophet) (3.2.3)
Requirement already satisfied: pytz>=2020.1 in d:\python\lib\site-packages (from pandas>=1.0.4->prophet) (2025.1)
Requirement already satisfied: tzdata>=2022.7 in d:\python\lib\site-packages (from pandas>=1.0.4->prophet) (2025.1)
Requirement already satisfied: colorama in d:\python\lib\site-packages (from tqdm>=4.36.1->prophet) (0.4.6)
Downloading prophet-1.2.1-py3-none-win_amd64.whl (12.1 MB)
   ---------------------------------------- 0.0/12.1 MB ? eta -:--:--
   ---- ----------------------------------- 1.3/12.1 MB 8.4 MB/s eta 0:00:02
   ------ --------------------------------- 1.8/12.1 MB 4.8 MB/s eta 0:00:03
   ----------- ---------------------------- 3.4/12.1 MB 5.6 MB/s eta 0:00:02
   -------------------- ------------------- 6.3/12.1 MB 7.7 MB/s eta 0:00:01
   ------------------------- -------------- 7.9/12.1 MB 8.1 MB/s eta 0:00:01
   ----------------------------- ---------- 8.9/12.1 MB 7.5 MB/s eta 0:00:01
   --------------------------------- ------ 10.0/12.1 MB 7.1 MB/s eta 0:00:01
   ------------------------------------ --- 11.0/12.1 MB 6.7 MB/s eta 0:00:01
   ---------------------------------------- 12.1/12.1 MB 6.4 MB/s eta 0:00:00
Downloading holidays-0.87-py3-none-any.whl (1.3 MB)
   ---------------------------------------- 0.0/1.3 MB ? eta -:--:--
   ----------------------- ---------------- 0.8/1.3 MB 4.2 MB/s eta 0:00:01
   ---------------------------------------- 1.3/1.3 MB 4.3 MB/s eta 0:00:00
Downloading cmdstanpy-1.3.0-py3-none-any.whl (99 kB)
Downloading stanio-0.5.1-py3-none-any.whl (8.1 kB)
Installing collected packages: stanio, holidays, cmdstanpy, prophet

   ---------- ----------------------------- 1/4 [holidays]
   ---------- ----------------------------- 1/4 [holidays]
   ---------- ----------------------------- 1/4 [holidays]
   ---------- ----------------------------- 1/4 [holidays]
   ---------- ----------------------------- 1/4 [holidays]
   ---------- ----------------------------- 1/4 [holidays]
   ---------- ----------------------------- 1/4 [holidays]
   ---------- ----------------------------- 1/4 [holidays]
   ---------- ----------------------------- 1/4 [holidays]
   ---------- ----------------------------- 1/4 [holidays]
   ------------------------------ --------- 3/4 [prophet]
   ------------------------------ --------- 3/4 [prophet]
   ------------------------------ --------- 3/4 [prophet]
   ------------------------------ --------- 3/4 [prophet]
   ------------------------------ --------- 3/4 [prophet]
   ------------------------------ --------- 3/4 [prophet]
   ---------------------------------------- 4/4 [prophet]

Successfully installed cmdstanpy-1.3.0 holidays-0.87 prophet-1.2.1 stanio-0.5.1

[notice] A new release of pip is available: 25.1.1 -> 25.3
[notice] To update, run: python.exe -m pip install --upgrade pip
```

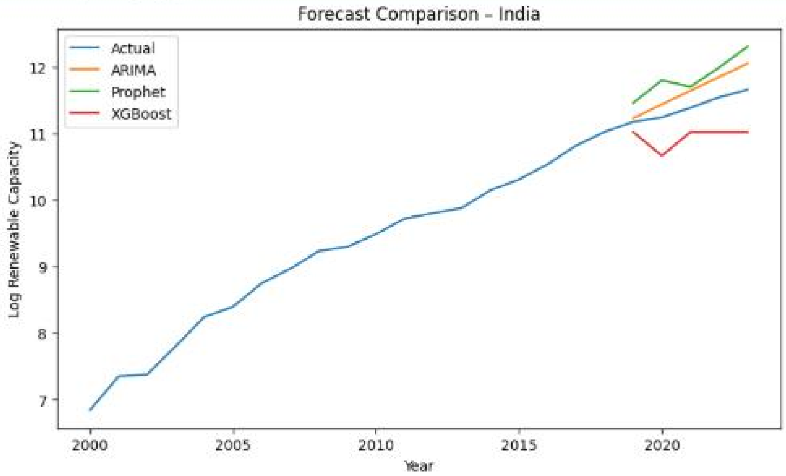
```python
import numpy as np
import pandas as pd
from statsmodels.tsa.arima.model import ARIMA
from prophet import Prophet
from sklearn.metrics import mean_absolute_error, mean_squared_error
from xgboost import XGBRegressor
import matplotlib.pyplot as plt

forecast_df = panel_df.reset_index()

forecast_df["log_total_re_capacity"] = np.log(forecast_df["total_re_capacity"] + 1)

countries_focus = [
    "Germany",
    "India",
    "United States of America (the)"
]

results = []

def evaluate_forecast(y_true, y_pred, model_name, country):
    rmse = np.sqrt(mean_squared_error(y_true, y_pred))
    mae = mean_absolute_error(y_true, y_pred)
    mape = np.mean(np.abs((y_true - y_pred) / (y_true + 1))) * 100
    results.append({
        "Country": country,
        "Model": model_name,
        "RMSE": rmse,
        "MAE": mae,
        "MAPE": mape
    })

for country in countries_focus:
    df_c = forecast_df[forecast_df["country"] == country].sort_values("year")

    y = df_c["log_total_re_capacity"].values
    years = df_c["year"].values

    split = int(len(y) * 0.8)
    y_train, y_test = y[:split], y[split:]
    years_test = years[split:]

    arima_model = ARIMA(y_train, order=(1,1,1)).fit()
    arima_forecast = arima_model.forecast(steps=len(y_test))
    evaluate_forecast(y_test, arima_forecast, "ARIMA", country)

    prophet_df = df_c[["year", "log_total_re_capacity"]].copy()
    prophet_df.columns = ["ds", "y"]
    prophet_df["ds"] = pd.to_datetime(prophet_df["ds"], format="%Y")

    prophet_train = prophet_df.iloc[:split]
    prophet_test = prophet_df.iloc[split:]

    prophet_model = Prophet()
    prophet_model.fit(prophet_train)
    future = prophet_model.make_future_dataframe(periods=len(prophet_test), freq="Y")
    prophet_forecast = prophet_model.predict(future)["yhat"].iloc[-len(prophet_test):].values
    evaluate_forecast(prophet_test["y"].values, prophet_forecast, "Prophet", country)

    features = [
        "log_gdp_per_capita",
        "log_population",
        "electricity_access_pct",
        "log_energy_use"
    ]

    X = df_c[features].values
    X_train, X_test = X[:split], X[split:]

    xgb = XGBRegressor(
        n_estimators=300,
        learning_rate=0.05,
        max_depth=4,
        subsample=0.8,
```

```
    prophet_model = Prophet()
    prophet_model.fit(prophet_train)
    future = prophet_model.make_future_dataframe(periods=len(prophet_test), freq="Y")
    prophet_forecast = prophet_model.predict(future)["yhat"].iloc[-len(prophet_test):].values
    evaluate_forecast(prophet_test["y"].values, prophet_forecast, "Prophet", country)

    features = [
        "log_gdp_per_capita",
        "log_population",
        "electricity_access_pct",
        "log_energy_use"
    ]

    X = df_c[features].values
    X_train, X_test = X[:split], X[split:]

    xgb = XGBRegressor(
        n_estimators=300,
        learning_rate=0.05,
        max_depth=4,
        subsample=0.8,
        colsample_bytree=0.8,
        random_state=42
    )

    xgb.fit(X_train, y_train)
    xgb_forecast = xgb.predict(X_test)
    evaluate_forecast(y_test, xgb_forecast, "XGBoost", country)

    plt.figure(figsize=(9,5))
    plt.plot(years, y, label="Actual")
    plt.plot(years_test, arima_forecast, label="ARIMA")
    plt.plot(years_test, prophet_forecast, label="Prophet")
    plt.plot(years_test, xgb_forecast, label="XGBoost")
    plt.title(f"Forecast Comparison - {country}")
    plt.xlabel("Year")
    plt.ylabel("Log Renewable Capacity")
    plt.legend()
    plt.show()

forecast_results = pd.DataFrame(results)
forecast_results
```

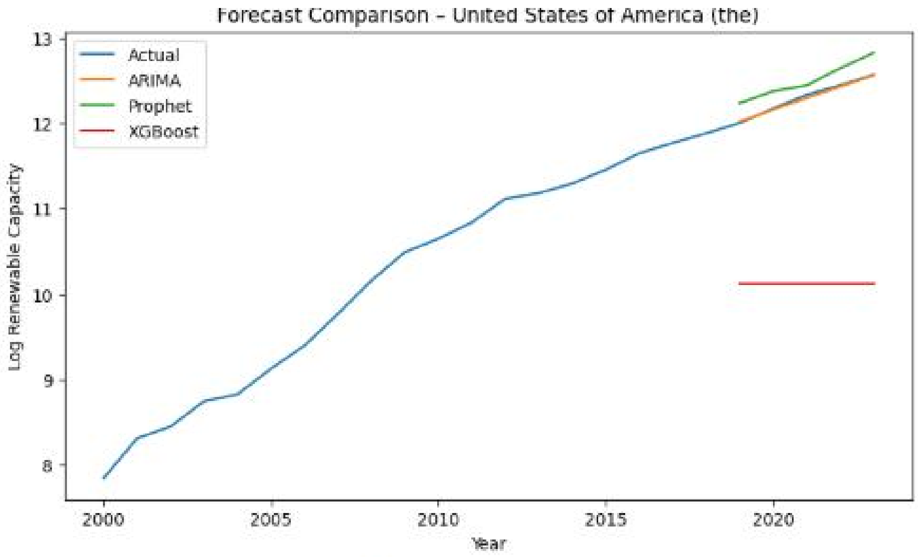
```
00:15:42 - cmdstanpy - INFO - Chain [1] start processing
00:15:42 - cmdstanpy - INFO - Chain [1] done processing
D:\Python\Lib\site-packages\prophet\forecaster.py:1872: FutureWarning: 'Y' is deprecated and will be removed in a future version, please use 'YE' instead.
  dates = pd.date_range(
```




```
00:15:43 - cmdstanpy - INFO - Chain [1] start processing
00:15:43 - cmdstanpy - INFO - Chain [1] done processing
D:\Python\Lib\site-packages\prophet\forecaster.py:1872: FutureWarning: 'Y' is deprecated and will be removed in a future version, please use 'YE' instead.
  dates = pd.date_range(
```



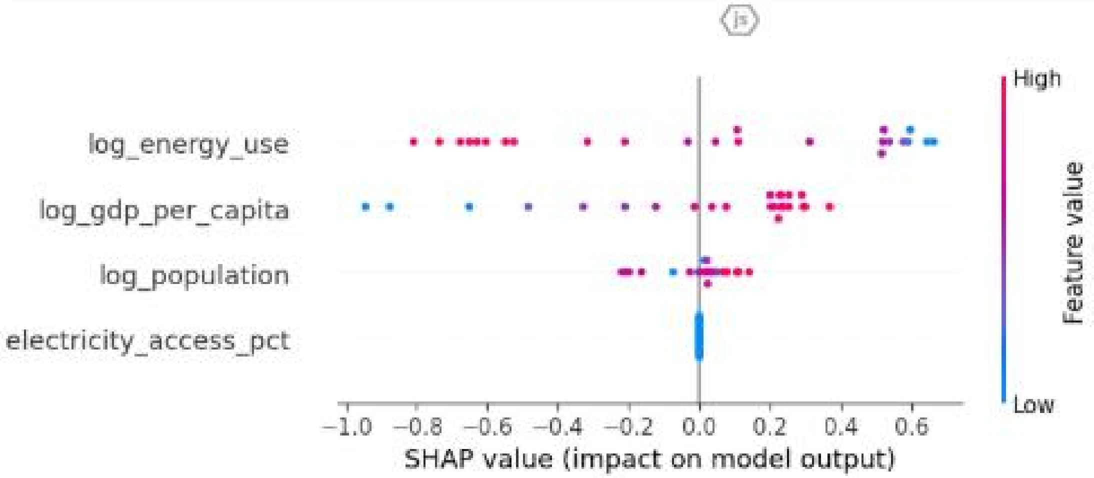

```
00:15:43 - cmdstanpy - INFO - Chain [1] start processing
00:15:52 - cmdstanpy - INFO - Chain [1] done processing
D:\Python\Lib\site-packages\prophet\forecaster.py:1872: FutureWarning: 'Y' is deprecated and will be removed in a future version, please use 'YE' instead.
  dates = pd.date_range(
```



	Country	Model	RMSE	MAE	MAPE
0	Germany	ARIMA	0.044782	0.026184	0.203905
1	Germany	Prophet	0.088320	0.079995	0.630211
2	Germany	XGBoost	0.220004	0.190229	1.489443
3	India	ARIMA	0.265870	0.240775	1.929395
4	India	Prophet	0.470325	0.449686	3.617595
5	India	XGBoost	0.483411	0.450979	3.623572
6	United States of America (the)	ARIMA	0.021133	0.018285	0.137751
7	United States of America (the)	Prophet	0.206350	0.200323	1.506337
8	United States of America (the)	XGBoost	2.191251	2.182215	16.388610

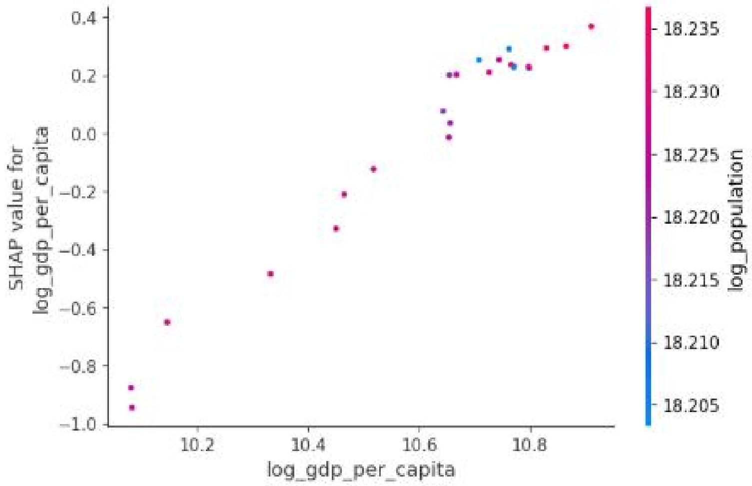
A9. Step 9: Explainability (SHAP analysis for XGBoost)

```python
import shap
import matplotlib.pyplot as plt
from xgboost import XGBRegressor

shap.initjs()

country_explain = "Germany"

df_shap = (
    panel_df
    .reset_index()
    .query("country == @country_explain")
    .sort_values("year")
)

features = [
    "log_gdp_per_capita",
    "log_population",
    "electricity_access_pct",
    "log_energy_use"
]

X_shap = df_shap[features].replace([np.inf, -np.inf], np.nan).dropna()
y_shap = np.log(df_shap.loc[X_shap.index, "total_re_capacity"] + 1)

xgb_model = XGBRegressor(
    n_estimators=300,
    learning_rate=0.05,
    max_depth=4,
    subsample=0.8,
    colsample_bytree=0.8,
    random_state=42
)

xgb_model.fit(X_shap, y_shap)

explainer = shap.Explainer(xgb_model, X_shap)
shap_values = explainer(X_shap)

shap.summary_plot(shap_values, X_shap)
```



[41]:

```
shap.dependence_plot(
    "log_gdp_per_capita",
    shap_values.values,
    X_shap
)
```



[42]:

```
panel_df.info()
panel_df.head()
```

```
<class 'pandas.core.frame.DataFrame'>
MultiIndex: 5064 entries, ('Afghanistan', np.int64(2000)) to ('Zimbabwe', np.int64(2023))
Data columns (total 17 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   solar_pv_capacity      5064 non-null   float64
 1   onshore_wind_capacity  5064 non-null   float64
 2   offshore_wind_capacity 5064 non-null   float64
 3   gdp_per_capita_usd     3710 non-null   float64
 4   population             3768 non-null   float64
 5   electricity_access_pct 3706 non-null   float64
 6   energy_use_per_capita  2874 non-null   float64
 7   total_wind_capacity    5064 non-null   float64
 8   total_re_capacity      5064 non-null   float64
 9   log_gdp_per_capita     3710 non-null   float64
 10  log_population         3768 non-null   float64
 11  log_energy_use         2874 non-null   float64
 12  imp_gdp                5064 non-null   int64
 13  imp_population         5064 non-null   int64
 14  imp_electricity        5064 non-null   int64
 15  imp_energy_use         5064 non-null   int64
 16  log_total_re_capacity  5064 non-null   float64
dtypes: float64(13), int64(4)
memory usage: 698.0+ KB
```

[42]:

solar pv capacity	onshore wind capacity	offshore wind capacity	gdp per capita usd	population	electricity access pct	energy use per capita

```
memory usage: 698.6+ KB
```

[42]:

country	year	solar pv capacity	onshore wind capacity	offshore wind capacity	gdp per capita usd	population	electricity access pct	energy use per capita
	2000	0.0	0.0	0.0	174.930991	20130327.0	4.4	NaN
	2001	0.0	0.0	0.0	138.706822	20284307.0	9.3	NaN
Afghanistan	2002	0.0	0.0	0.0	178.954088	21378117.0	14.1	NaN
	2003	0.0	0.0	0.0	198.871116	22733049.0	19.0	NaN
	2004	0.0	0.0	0.0	221.763654	23560654.0	23.8	NaN

[43]:

```
desc_stats
```

[43]:

	solar pv capacity	onshore wind capacity	offshore wind capacity	total re capacity	gdp per capita usd	population	electricity access pct	energy use per
count	5064.000000	5064.000000	5064.000000	5.064000e+03	3710.000000	3.768000e+03	3706.000000	2874.
mean	70.661291	1513.442289	1290.458118	2.874562e+03	13937.068720	3.670099e+07	81.140340	2378.
std	956.433856	12513.188270	13584.681056	2.626029e+04	19676.038081	1.498006e+08	29.110452	2964.
min	0.000000	0.000000	0.000000	0.000000e+00	109.593814	9.544000e+03	0.800000	9.
25%	0.000000	0.000000	0.000000	0.000000e+00	1580.563837	1.219356e+06	70.325000	560.
50%	0.000000	0.000000	0.220000	2.000000e+00	5141.779434	6.825680e+06	98.900000	1381.
75%	0.000000	41.900000	23.000000	1.178200e+02	18629.317824	2.073615e+07	100.000000	2856.
max	37290.000000	404050.000000	608920.000000	1.050260e+06	134965.815442	1.438070e+09	100.000000	21557.

[44]:

```
global_trends
```

[44]:

	year	solar pv capacity	onshore wind capacity	offshore wind capacity	total re capacity
0	2000	66.95	16881.33	755.69	17703.97
1	2001	75.95	23872.38	1022.80	24971.13
2	2002	239.95	30474.36	1341.02	32055.33
3	2003	510.55	38142.98	1847.47	40501.00
4	2004	595.75	47057.51	2915.31	50568.57
5	2005	685.75	57770.35	4390.35	62846.45
6	2006	883.75	72252.81	5911.64	79048.20
7	2007	1094.25	90411.65	8265.92	99771.82
8	2008	1442.45	114019.74	14444.86	129907.05
9	2009	2133.55	147956.61	22483.40	172573.56
10	2010	3055.35	177711.64	39732.76	220499.75
11	2011	3786.61	216246.13	70189.34	290222.08
12	2012	5344.75	261361.27	99017.28	365723.30
13	2013	7182.05	292277.12	134040.27	433499.44

[45]:

```
top_countries
```

[45]:

```
country
China                                                   1050260.00
United States of America (the)                           285966.90
Germany                                                  144368.00
India                                                    115498.56
Japan                                                     92437.04
Brazil                                                    67049.33
Spain                                                     60448.11
United Kingdom of Great Britain and Northern Ireland (the)  46438.21
Australia                                                 43579.94
Italy                                                     41658.70
France                                                    40157.34
Netherlands (Kingdom of the)                              32008.76
Poland                                                    25770.85
Republic of Korea (the)                                   25692.94
Viet Nam                                                  24464.42
Name: total_re_capacity, dtype: float64
```

[46]:

```
corr_matrix
```

[46]:

	log_gdp_per_capita	log_population	log_energy_use	electricity_access_pct	total_re_capacity
log_gdp_per_capita	1.000000	-0.253789	0.845150	0.728215	0.080116
log_population	-0.253789	1.000000	-0.198849	-0.165896	0.185957
log_energy_use	0.845150	-0.198849	1.000000	0.689415	0.068859
electricity_access_pct	0.728215	-0.165896	0.689415	1.000000	0.066821
total_re_capacity	0.080116	0.185957	0.068859	0.066821	1.000000

[47]:

```
pooled_ols.summary()
```

[47]:

OLS Regression Results

Dep. Variable:	log_total_re_capacity	**R-squared:**	0.595
Model:	OLS	**Adj. R-squared:**	0.595
Method:	Least Squares	**F-statistic:**	1387.
Date:	Thu, 18 Dec 2025	**Prob (F-statistic):**	0.00
Time:	00:32:18	**Log-Likelihood:**	-6225.9
No. Observations:	2850	**AIC:**	1.246e+04

[47]:

```
pooled_ols.summary()
```

[47]:

OLS Regression Results

Dep. Variable:	log_total_re_capacity	**R-squared:**	0.595
Model:	OLS	**Adj. R-squared:**	0.595
Method:	Least Squares	**F-statistic:**	1387.
Date:	Thu, 18 Dec 2025	**Prob (F-statistic):**	0.00
Time:	00:32:18	**Log-Likelihood:**	-6225.9
No. Observations:	2850	**AIC:**	1.246e+04
Df Residuals:	2845	**BIC:**	1.249e+04
Df Model:	4		
Covariance Type:	HC3		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-24.0885	0.490	-49.157	0.000	-25.049	-23.128
log_gdp_per_capita	2.0757	0.048	43.093	0.000	1.981	2.170
log_population	1.0328	0.025	41.512	0.000	0.984	1.082
electricity_access_pct	0.0135	0.002	6.815	0.000	0.010	0.017
log_energy_use	-1.1472	0.066	-17.466	0.000	-1.276	-1.018

Omnibus:	106.431	**Durbin-Watson:**	0.173
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	69.109
Skew:	-0.256	**Prob(JB):**	9.84e-16
Kurtosis:	2.435	**Cond. No.**	1.20e+03

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)
[2] The condition number is large, 1.2e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

[48]:

```
fe_model.summary
```

[48]:

PanelOLS Estimation Summary

Dep. Variable:	log_total_re_capacity	**R-squared:**	0.4716
Estimator:	PanelOLS	**R-squared (Between):**	-2.6336
No. Observations:	2850	**R-squared (Within):**	0.4716
Date:	Wed, Dec 17 2025	**R-squared (Overall):**	-0.8098
Time:	23:53:45	**Log-likelihood**	-5153.3
Cov. Estimator:	Clustered		

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)
[2] The condition number is large, 1.2e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

48]:

```
fe_model.summary
```

48]:

PanelOLS Estimation Summary

Dep. Variable:	log_total_re_capacity	**R-squared:**	0.4716
Estimator:	PanelOLS	**R-squared (Between):**	-2.6336
No. Observations:	2850	**R-squared (Within):**	0.4716
Date:	Wed, Dec 17 2025	**R-squared (Overall):**	-0.8098
Time:	23:53:45	**Log-likelihood**	-5153.3
Cov. Estimator:	Clustered		
		F-statistic:	605.12
Entities:	134	**P-value**	0.0000
Avg Obs:	21.269	**Distribution:**	F(4,2712)
Min Obs:	6.0000		
Max Obs:	24.000	**F-statistic (robust):**	68.808
		P-value	0.0000
Time periods:	24	**Distribution:**	F(4,2712)
Avg Obs:	118.75		
Min Obs:	113.00		
Max Obs:	133.00		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	-67.883	14.216	-4.7752	0.0000	-95.758	-40.008
log_gdp_per_capita	2.3408	0.2824	8.2882	0.0000	1.7870	2.8946
log_population	3.4177	0.8824	3.8733	0.0001	1.6875	5.1479
electricity_access_pct	0.0233	0.0115	2.0191	0.0436	0.0007	0.0459
log_energy_use	-0.8259	0.7507	-1.1001	0.2714	-2.2980	0.6462

F-test for Poolability: 22.896
P-value: 0.0000
Distribution: F(133,2712)

Included effects: Entity

49]:

```
re_model.summary
```

```
re_model.summary
```

[49]:

RandomEffects Estimation Summary

Dep. Variable:	log_total_re_capacity	**R-squared:**	0.4643
Estimator:	RandomEffects	**R-squared (Between):**	0.6271
No. Observations:	2850	**R-squared (Within):**	0.4558
Date:	Wed, Dec 17 2025	**R-squared (Overall):**	0.5589
Time:	23:53:45	**Log-likelihood**	-5279.6
Cov. Estimator:	Clustered		
		F-statistic:	616.54
Entities:	134	**P-value**	0.0000
Avg Obs:	21.269	**Distribution:**	F(4,2845)
Min Obs:	6.0000		
Max Obs:	24.000	**F-statistic (robust):**	120.06
		P-value	0.0000
Time periods:	24	**Distribution:**	F(4,2845)
Avg Obs:	118.75		
Min Obs:	113.00		
Max Obs:	133.00		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	-27.892	1.8170	-15.351	0.0000	-31.455	-24.330
log_gdp_per_capita	2.5584	0.2155	11.871	0.0000	2.1358	2.9810
log_population	1.1869	0.0899	13.199	0.0000	1.0106	1.3632
electricity_access_pct	0.0298	0.0080	3.7110	0.0002	0.0141	0.0456
log_energy_use	-1.7251	0.2786	-6.1923	0.0000	-2.2713	-1.1788

[50]:

```
comparison
```

[50]:

Model Comparison

	Fixed Effects	Random Effects
Dep. Variable	log_total_re_capacity	log_total_re_capacity
Estimator	PanelOLS	RandomEffects
No. Observations	2850	2850
Cov. Est.	Clustered	Clustered
R-squared	0.4716	0.4643
R-Squared (Within)	0.4716	0.4558
R-Squared (Between)	-2.6336	0.6271
R-Squared (Overall)	-0.8098	0.5589
F-statistic	605.12	616.54
P-value (F-stat)	0.0000	0.0000
=====================	=====================	=====================
const	-67.883	-27.892
	(-4.7752)	(-15.351)
log_gdp_per_capita	2.3408	2.5584
	(8.2882)	(11.871)
log_population	3.4177	1.1869

Model Comparison

	Fixed Effects	Random Effects
Dep. Variable	log_total_re_capacity	log_total_re_capacity
Estimator	PanelOLS	RandomEffects
No. Observations	2850	2850
Cov. Est.	Clustered	Clustered
R-squared	0.4716	0.4643
R-Squared (Within)	0.4716	0.4558
R-Squared (Between)	-2.6336	0.6271
R-Squared (Overall)	-0.8098	0.5589
F-statistic	605.12	616.54
P-value (F-stat)	0.0000	0.0000
========================	======================	======================
const	-67.883	-27.892
	(-4.7752)	(-15.351)
log_gdp_per_capita	2.3408	2.5584
	(8.2882)	(11.871)
log_population	3.4177	1.1869
	(3.8733)	(13.199)
electricity_access_pct	0.0233	0.0298
	(2.0191)	(3.7110)
log_energy_use	-0.8259	-1.7251
	(-1.1001)	(-6.1923)
=========================	=========================	=========================
Effects	Entity	

T-stats reported in parentheses
id: 0x1f3486a3590

```
vif_data
```

	Variable	VIF
0	const	176.675914
1	log_gdp_per_capita	3.952011
2	log_population	1.075614
3	electricity_access_pct	2.084202
4	log_energy_use	3.746862

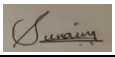
Declaration of Authenticity

I hereby declare that this master's thesis has been composed by me and that the work contained herein is my own, except where explicitly stated otherwise. All sources and references used in the preparation of this thesis have been appropriately acknowledged and cited in accordance with academic standards. This thesis has not been submitted, in whole or in part, for the award of any other degree or qualification at this or any other institution.

I confirm that this thesis complies with the academic integrity regulations of IU International University of Applied Sciences.

Place: Berlin

Date: 20th January 2026

Signature: 

Name: Sunaina Manjunath