



UGANDA CHRISTIAN
UNIVERSITY

A Centre of Excellence in the Heart of Africa

Department of Computing & Technology

*Machine Learning Forecasting for Electricity Demand Using
Climate and Socioeconomic Data in Sub-Saharan Africa*

By

LORDIN LUTALO

Bachelor of Information Technology & Computing

Reg No: *J25M19/004*, Student No: *J25M19/004*

E-mail: *B31327@students.ucu.ac.ug*, Phone No: *+256 706751625*

A report submitted to the Faculty of Engineering Design &
Technology for the study leading to the Degree in partial fulfillment
of the requirements for the award of **Master of Science in Data
Science & Analytics** of Uganda Christian University

September 2026

Declaration

I, *LORDIN LUTALO*, declare to the best of my knowledge that the work in this report is original and has not been submitted before to any university/institution. It is the result of my effort.

Signature:  _____

Date: 15/09/2026

Ethical Generative AI Usage Declaration

Students may use generative artificial intelligence (AI) ethically as a “study buddy” or assistant rather than a replacement for critical thinking. This includes citing AI use and fact-checking all outputs to maintain academic integrity.

For purposes of this assignment, the following shall guidelines shall apply.

- i. Generative AI may be used for brainstorming, and generating ideas to improve your work. You may also use AI for literature search, and summarizing journal papers for quick analysis and assessing relevance of publications. However, once the relevant publications have been identified, a student is expected to fully review the selected publications, identifying problems addressed, methods used, gaps identified and recommendations made.
- ii. Where AI has been used as guided in (i) above, full disclosure of how it has been used should be made in the appendix **B**, including how AI was used, the AI tools (including mode versions), and a compilation on major AI prompts used.
- iii. NO GENERATIVE AI CONTENT IS ALLOWED IN THE FINAL SUBMISSION. The final submission SHOULD be your original work.

A. Declaring the non-use of AI for content generation

I, *LORDIN LUTALO*, declare that no content generated by AI technologies has been used in this report.

Signature:  Date: 15/09/2026

B. Acknowledging the use of AI as a study buddy

Please refer to Appendix **B** for the complete declaration of AI usage for generating ideas, brainstorming, literature search, and summarizing only to improve your work.

Approval

This report has been submitted for examination with approval of the following advisors/supervisor.

Advisor: **Dr. Terhemba Michael Ahile**
Department of Computing & Technology
Faculty of Engineering Design & Technology
Uganda Christian University

Signature:  _____

Date: 24/09/2026

Co-Advisor: **Mr. Brian Senabulya**
Department of Computing & Technology
Faculty of Engineering Design & Technology
Uganda Christian University

Signature:  _____

Date: 24/09/2026

Abstract

Electricity demand forecasting is important for energy planning, resource allocation, and infrastructure development. In Sub-Saharan Africa, forecasting electricity demand is particularly difficult because electricity usage is driven by rapid population growth, economic changes, urbanisation, and climate variability, while data quality and availability remain inconsistent across countries. Although previous studies showed that climate and socioeconomic variables are associated with electricity demand, many focused on short-term forecasting, single-country analysis, or datasets from developed regions.

This study examined how far integrated climate and socioeconomic data improves annual electricity demand forecasting across 47 Sub-Saharan African countries. Country-level panel time-series data was collected from publicly available sources including the World Bank, Our World in Data, and the ERA5 climate reanalysis dataset, covering the period from 2000 to 2021. The methodology included data integration and preprocessing, normality testing, temporal feature engineering, a two-stage feature selection process, and the training and comparison of six forecasting models: Linear Regression, Fixed Effects Panel Regression, Dynamic Fixed Effects Panel Regression, Random Forest, XGBoost, and LightGBM. Model performance was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination (R^2). A decision-support dashboard was also developed using Streamlit to visualise electricity demand forecasts, trends, and key demand drivers.

The results showed that lagged electricity demand was by far the strongest predictor of future demand, with the one-year lag achieving a near-perfect Pearson correlation of 0.999 with the target variable. Among structural climate and socioeconomic predictors, population size and its three-year rolling mean were the strongest predictors of electricity demand, followed by the three-year rolling mean of temperature, the interaction between population and electricity access, and the temperature-urbanisation interaction. Linear Regression on the full multivariate feature set achieved the best overall test set performance, with a log-scale MAPE of 6.01% and an R^2 of 0.996 (note: all metrics are on the log-transformed demand scale, not in original TWh units), and was selected as the final model for the dashboard. The study confirms that annual electricity demand in Sub-Saharan Africa is extremely persistent, lagged demand provides most of the predictive power in the model. Climate and socioeconomic variables contain substantial structural information about demand patterns when historical demand is unavailable, but contribute little additional one-year-ahead predictive accuracy once lagged demand is already included. This distinction between predictive accuracy and structural explanation is the core finding of the study.

Dedication

I dedicate this report to my family for their continuous support and encouragement throughout my academic journey. I also dedicate this work to all individuals and institutions working towards improving energy access and sustainable development in Sub-Saharan Africa.

Contents

Declaration	i
Declaration	ii
Approval	iii
Abstract	iv
Dedication	v
Acknowledgments	xi
1 Introduction	1
1.1 Background of the Study	2
1.2 Problem Statement	2
1.3 Research Objectives	3
1.3.1 General Objective	3
1.3.2 Specific Objective 1	3
1.3.3 Specific Objective 2	3
1.3.4 Specific Objective 3	3
1.4 Research Questions	3
1.4.1 Research Question 1	3
1.4.2 Research Question 2	3
1.4.3 Research Question 3	4
1.5 Significance of the Study	4
1.6 Scope of the Study	4
1.7 Limitations of the Study	5
1.8 Definition of Key Terms	5
2 Literature Review	8
2.1 Introduction	8
2.2 Types of Electricity Demand Forecasting	8
2.3 Electricity Demand Forecasting Approaches	9
2.3.1 Traditional Statistical Methods	9
2.3.2 Machine Learning Approaches	9
2.4 Integration of Climate and Socioeconomic Variables	10
2.4.1 Climate Variables	10
2.4.2 Socioeconomic Variables	11
2.5 Challenges in Electricity Demand Forecasting	11
2.5.1 Data Quality and Integration	11

2.5.2	Model Overfitting and Generalisation	11
2.5.3	Feature Selection	12
2.5.4	Geographic Focus of Existing Research	12
2.6	Comparative Analysis of Existing Studies	12
2.7	Critical Analysis of Existing Studies	13
2.8	Research Gap	14
2.9	Chapter Summary	15
3	Methodology	16
3.1	Introduction	16
3.2	Research Approach	16
3.3	Research Design	17
3.3.1	Forecasting Strategy	17
3.3.2	Analytical Framework	17
3.4	Data Sources and Dataset Description	18
3.4.1	Data Sources	18
3.4.2	Dataset Description	18
3.4.3	Variable Selection and Justification	18
3.5	Data Integration and Preprocessing	20
3.5.1	Data Integration	20
3.5.2	Outlier Treatment	21
3.5.3	Normality Testing	21
3.5.4	Missing Value Treatment	21
3.5.5	Scaling	21
3.6	Feature Engineering	22
3.6.1	Log Transformation	22
3.6.2	Lag Features	23
3.6.3	Growth Rate Features	23
3.6.4	Interaction Features	23
3.6.5	Trend Feature	24
3.6.6	Climate Anomaly Features	24
3.6.7	Rolling Average Features	24
3.6.8	Temperature Variability	25
3.6.9	Engineered Features Summary	26
3.7	Feature Selection	26
3.7.1	Stage 1: Temporal Feature Evaluation	27
3.7.2	Stage 2: Exogenous Feature Selection	27
3.7.3	Three Feature Sets	29
3.8	Model Training	30
3.8.1	Linear Regression	30
3.8.2	Fixed Effects Panel Regression	30
3.8.3	Dynamic Fixed Effects Panel Regression	31
3.8.4	Random Forest	31
3.8.5	XGBoost	31
3.8.6	LightGBM	31
3.9	Hyperparameter Tuning	32
3.10	Model Evaluation	32
3.11	Training, Validation, and Data Integrity	33
3.11.1	Train-Test Split	33
3.11.2	Walk-Forward Validation	34

3.11.3 Data Leakage Prevention	34
3.12 Visualization and Interpretation	35
3.13 Methodological Framework	36
3.14 Ethical Considerations	37
3.14.1 Data Integrity and Responsible Use	37
3.14.2 Fair Representation of Countries	37
3.14.3 Transparency and Explainability	38
3.14.4 Avoidance of Misuse	38
3.14.5 Data Sovereignty and Publication of Results	38
3.15 Methodological Constraints	38
4 Results	40
4.1 Introduction	40
4.2 Dataset Integration and Descriptive Statistics	40
4.2.1 Demand Concentration	41
4.3 Exploratory Data Analysis	42
4.3.1 Temporal Trends	42
4.3.2 Missing Values by Year	45
4.4 Preprocessing Results	46
4.4.1 Normality Test Results	46
4.4.2 Outlier Analysis	47
4.4.3 Missing Value Treatment	48
4.5 Feature Engineering Results	49
4.5.1 Log Transformation	49
4.5.2 Feature Engineering Summary	50
4.6 Feature Selection Results	50
4.6.1 Stage 1: Full Feature Evaluation and Lag Consistency Check	50
4.6.2 Stage 2: Exogenous Feature Selection	55
4.6.3 Consolidated Exogenous Feature Ranking	59
4.6.4 Final Feature Sets	60
4.7 Model Performance Results	60
4.7.1 Hyperparameter Tuning Results	60
4.7.2 Walk-Forward Validation Results	61
4.7.3 Best Model Per Feature Set	62
4.7.4 Final Test Set Results	62
4.7.5 Three-Way Feature Set Comparison	62
4.8 Permutation Importance Results	63
4.9 Final Model Selection	64
4.10 Decision-Support Dashboard	65
5 Discussion	70
5.1 Introduction	70
5.2 Why Lag Features Dominate Annual Electricity Demand Forecasting	70
5.3 Why Population Is the Most Important Structural Driver	71
5.4 Why GDP Contributes Less Once Lags Are Included	72
5.5 Why Temperature Shows a Negative Relationship with Demand	73
5.6 What the Exogenous-Only Results Tell Us About Structural Demand Drivers	74
5.7 Why Linear Regression Outperformed More Complex Models	75
5.8 Why Model C Was Selected Despite Near-Identical Performance to Model A	76
5.9 What the Climate and Socioeconomic Variables Contribute	77

5.10	What Permutation Importance Confirmed Post-Modelling	78
5.11	Comparison with the Existing Literature	79
5.12	Significance of the Findings	79
5.13	Implications for Energy Planning and Policy in Sub-Saharan Africa	80
5.14	Limitations and Their Implications	81
5.15	Chapter Summary	82
6	Conclusion and Recommendation	83
6.1	Introduction	83
6.2	Summary of the Study	83
6.3	Answers to the Research Questions	84
6.3.1	Research Question 1: Which engineered climate, socioeconomic and temporal features are most strongly associated with annual electricity demand in Sub-Saharan Africa?	84
6.3.2	Research Question 2: How much does forecasting accuracy improve when climate and socioeconomic variables are added to a model using past demand alone, and which modelling approach performs best on each feature set?	85
6.3.3	Research Question 3: How can a decision-support dashboard present electricity demand forecasts, trends and demand drivers in a form that energy planners and policy analysts can use directly?	85
6.4	Contributions of the Study	86
6.4.1	Integrated Multi-Country Dataset	86
6.4.2	Two-Stage Feature Selection Methodology	86
6.4.3	Structured Three-Way Model Comparison	86
6.4.4	Contextual Insights for Sub-Saharan Africa	86
6.4.5	Decision-Support Dashboard	87
6.5	Recommendations for Future Research	87
6.5.1	Use of Higher-Frequency Data	87
6.5.2	Inclusion of Electricity Prices	87
6.5.3	Inclusion of Drought Indices	87
6.5.4	Country-Specific Models	87
6.5.5	Exploration of Deep Learning Methods	88
6.5.6	Extension to Future Projections	88
6.5.7	Integration of Policy and Institutional Variables	88
6.6	Final Remarks	88
	REFERENCES	92
	APPENDICES	92
A	Research Project timelines	93
A.1	Project Management Plan	93
A.2	Project Phases and Tasks	94
B	Declaration of AI use	95
C	List of 20 Major Journal Publications reviewed	96
D	Research alignment with national (NDPIV) and Global (SDGs) agenda	98
D.1	Alignment with the National Development Agenda	98

D.1.1	Alignment with MSc Data Science and Analytics	98
D.1.2	Alignment with NDPIV	98
D.1.3	Alignment with the Uganda Vision 2040	99
D.1.4	Alignment with Other National Agendas	99
D.2	Alignment with Global Agenda	99
D.2.1	Alignment with Sustainable Development Goals (SDGs)	99
D.3	Alignment with Global Energy Transition Initiatives	100
D.4	Alignment with Climate and Sustainability Frameworks	100
E	References to Datasets and Tools Used	101
E.1	Datasets	101
E.2	Software and Tools	101
F	Walk-Forward Validation Detailed Results	103
F.1	Set A: Lag-Only Feature Set	103
F.2	Set B: Exogenous-Only Feature Set	104
F.3	Set C: Full Multivariate Feature Set	104
G	Research poster layout	105
H	Research Ethics Committee (REC) Approval	108
I	Student - supervisor meeting/supervision plan	110

Acknowledgments

I would like to express my sincere gratitude to Dr. Tahemba Ahile and Mr. Brian Senabulya for their guidance and support throughout the development of this report.

I also thank the research community for sharing useful resources. I used public datasets and open-source tools that helped me a lot in completing this study.

Chapter 1

Introduction

Executive Summary

This study examined what climate and socioeconomic data adds to annual electricity demand forecasting in Sub-Saharan Africa. Data from 47 countries was used, covering the years 2000 to 2021. Seven different datasets were pulled together from sources like Our World in Data, the World Bank, and the ERA5 climate dataset from the World Bank Climate Change Knowledge Portal. After putting all this data together, 27 features were created to help the models predict electricity demand. These features included things like past demand values (lag features), how fast key variables were changing (growth rate features), how much the climate deviated from normal (anomaly features), smoothed averages (rolling average features), and combined variables (interaction features).

Six forecasting models were then trained and tested using three groups of features: one group with only past demand values (Model A), one with only climate and socioeconomic variables (Model B), and one that combined everything together (Model C). The models were trained and validated on data from 2000 to 2017, and then tested on the period 2018 to 2021, which the models had never seen before. The best-performing model was Linear Regression using all features together (Model C), which gave a log-scale MAPE of 6.01%, an RMSE of 0.068, and an R^2 of 0.996 on the test set. All metrics were computed on the log-transformed demand values rather than on original terawatt-hours. The very high R^2 mostly reflects how strongly electricity demand carries over from one year to the next, rather than showing that the climate and socioeconomic variables were doing the explaining. The permutation importance analysis later confirmed this. A Streamlit dashboard was also built so that users can explore the forecasting results, electricity demand trends, and the key factors associated with demand in this context.

1.1 Background of the Study

Electricity underpins almost every part of modern life. Hospitals, schools, factories, transport systems and communication networks all depend on a steady supply. Global electricity use rose steadily through the study period, pushed up by economic growth and by the spread of technologies that run on electricity [1].

The rise was sharper in Sub-Saharan Africa, driven by fast population growth, the spread of cities, and the extension of the grid to new communities. Even so, supply remained unreliable across much of the region and a large share of the population still had no dependable access to electricity [2]. Businesses in the region lost an estimated 8 per cent of their annual revenue to power disruptions, which shows how directly the shortfall reaches the wider economy [2].

At national level the planning burden falls on individual ministries and utilities working with thin data. Uganda illustrates the position. The Energy Policy for Uganda 2023 projects demand of 7,664 GWh by 2027 against a least-cost generation plan that raises capacity from 872 MW in 2018 to 3,127 MW in 2027 [3]. Figures of this kind determine which plants and transmission lines are financed, yet they are usually built from demand trends and a small number of economic assumptions rather than from a tested account of which drivers actually matter [4].

Better forecasting also serves commitments the region has already made. Sustainable Development Goal 7 commits countries to affordable and reliable modern energy for all by 2030, and Goal 13 commits them to act on climate change [5]. The African Union’s Agenda 2063 treats energy infrastructure as a foundation for industrial growth [6]. Each of these depends on knowing how much electricity will be needed and where, because generation and grid projects take years to build and are expensive to reverse.

Existing research offers limited help at this level. Most studies are short-term, cover a single country, or are set in developed regions where high-frequency data is easy to obtain [7, 8], and combining electricity, climate and socioeconomic datasets across many countries remains difficult [9]. This study therefore used integrated climate and socioeconomic data from 47 Sub-Saharan African countries covering the years 2000 to 2021, with temporal feature engineering applied to support both prediction and interpretation.

1.2 Problem Statement

Throughout the study period, electricity demand across Sub-Saharan Africa kept rising steadily. This was driven by fast population growth, the spread of cities, more industrial activity, and the ongoing effort to connect more people to the grid [1, 2]. Even so, more than 600 million people in the region still had no access to electricity [2], and many countries kept facing unreliable supply, regular outages, and infrastructure that was not well developed. The World Bank estimated that businesses in the region lost around 8% of their yearly revenue because of power supply disruptions [2], while poor demand forecasting added to the problem by leading to overproduction, underproduction, weak resource allocation, and higher running costs [4].

This is why forecasting electricity demand accurately matters so much for planning generation, transmission, distribution, and infrastructure investment [4]. The problem is that this is not an easy task, because electricity demand is influenced by a lot of factors that interact with each other, including climate conditions, economic performance, population dynamics, urbanisation, and electricity access [10, 11]. Older forecasting methods depended mostly on past demand and often could not capture these complicated, nonlinear relationships [12]. Some more recent

studies did bring in climate and socioeconomic variables, but most of them concentrated on short-term forecasting, single countries, or developed regions [7, 8, 13], and very few built integrated models for several Sub-Saharan African countries using country-level data. On top of that, the fact that electricity, climate, and socioeconomic datasets come from different sources, in different formats, and with missing values made integration and forecasting even harder [9]. All of this pointed to a clear need for a machine learning forecasting model that brought together climate and socioeconomic variables using country-level historical data, in order to improve electricity demand forecasting in Sub-Saharan Africa.

1.3 Research Objectives

1.3.1 General Objective

To establish how far integrated climate and socioeconomic data can improve annual electricity demand forecasting in Sub-Saharan Africa, and to make the resulting forecasts and demand drivers available to energy planners through a decision-support tool.

1.3.2 Specific Objective 1

To determine which engineered climate, socioeconomic and temporal features are most strongly associated with annual electricity demand across Sub-Saharan African countries.

1.3.3 Specific Objective 2

To determine how much forecasting accuracy is gained by adding climate and socioeconomic variables to a model built on past demand alone, and which modelling approach performs best in each case.

1.3.4 Specific Objective 3

To present the forecasts, demand trends and demand drivers through a decision-support dashboard that energy planners and policy analysts can use without a machine learning background.

1.4 Research Questions

1.4.1 Research Question 1

Which engineered climate, socioeconomic and temporal features are most strongly associated with annual electricity demand in Sub-Saharan Africa?

1.4.2 Research Question 2

How much does forecasting accuracy improve when climate and socioeconomic variables are added to a model using past demand alone, and which modelling approach performs best on each feature set?

1.4.3 Research Question 3

How can a decision-support dashboard present electricity demand forecasts, trends and demand drivers in a form that energy planners and policy analysts can use directly?

1.5 Significance of the Study

One of the main contributions of this study was bringing together a wide range of data sources and machine learning techniques to forecast electricity demand across Sub-Saharan Africa. Forecasting electricity demand accurately has a lot of practical value. It helps governments and energy companies plan more wisely, avoid wasting resources, and build the right infrastructure in the right places.

At a global level, poor energy planning can be very costly. When too much or too little electricity is generated, money is wasted and energy systems become unstable. The IEA has estimated that improvements in energy efficiency alone could reduce global energy costs by somewhere between 10 and 15% [14]. For Sub-Saharan Africa in particular, where many countries are already dealing with electricity shortages and ageing infrastructure, getting forecasting right is even more critical [2].

This study also stood out because it did not just rely on past electricity demand data. It brought in climate variables and socioeconomic indicators and tried to understand how all of these factors are associated with electricity demand patterns. Research has shown that things like temperature and rainfall affect how much electricity people use, while factors like economic growth, population increases, and how many people are moving to cities determine how demand trends change over the long term [10, 11]. The feature engineering work done in this study, which included creating lag features, growth rate features, anomaly features, and rolling averages, was aimed at capturing all of these effects as fully as possible [12, 10].

Another important aspect of this study was that it covered multiple countries at the same time. Most previous studies in Sub-Saharan Africa focused on just one country or one sector [7, 8]. By looking at 47 countries together, this study produced findings that are more broadly applicable and contribute more meaningfully to what is known about electricity demand in developing regions.

As set out in the background, the study supports SDG 7 and SDG 13 by strengthening the evidence behind the demand projections that underpin generation and grid investment decisions [5].

The results and the dashboard built as part of this study are intended to be useful not just to researchers, but also to energy planners, policymakers, and development organisations who are working to improve electricity access and planning in the region.

1.6 Scope of the Study

This study focused on forecasting annual electricity demand in Sub-Saharan Africa using country-level panel time-series data. The data was collected from a number of well-known public sources, including the World Bank, Our World in Data, and the ERA5 climate reanalysis dataset, which was accessed through the World Bank Climate Change Knowledge Portal [1, 2].

The study period ran from 2000 to 2021. This particular time period was chosen because it offered the best balance of data availability⁴ and consistency across all the countries and

variables included in the study. South Sudan was the only country removed from the analysis. Because it only became an independent country in 2011, it had too many years of missing data to be included without risking unreliable imputations.

In total, six forecasting models were used in the study: Linear Regression, Fixed Effects Panel Regression, Dynamic Fixed Effects Panel Regression, Random Forest, XGBoost, and LightGBM [15]. Model performance was assessed using four widely used forecasting metrics: MAE, RMSE, MAPE, and R^2 [12, 16].

The study also produced an interactive dashboard using Streamlit, which allows users to explore the forecasting results and electricity demand trends. It is important to note that the study did not attempt to do real-time forecasting, and the models were not deployed directly into any national energy systems.

1.7 Limitations of the Study

Like most research studies, this one had a number of limitations. The first and probably most obvious one is that the data used was not perfect. Some countries had missing values in their records, and the way data was collected and reported varied quite a bit from one country to another. To deal with this, missing values were filled in using linear interpolation within each country's own data, and for gaps at the edges of the time series where interpolation could not work, forward fill was applied instead, using only past values so that no future information leaked backward into earlier years [17].

A second limitation is that the data was annual. This means the study looked at electricity demand year by year, rather than month by month or day by day. Annual data is good for looking at long-term trends, but it does miss things like seasonal peaks in demand or the effect of a particularly hot or cold week. Monthly or daily data would have been more informative, but it was simply not available for all 47 countries across the full 22-year period [7].

There were also a few variables that would have been helpful to include but could not be because the data was not consistently available. Things like electricity prices, government energy policies, and how consumers actually behave when it comes to electricity use all play a role in demand, but tracking these across 47 countries over 22 years was not feasible with publicly available data.

Drought data was another thing the study would have benefited from including. Droughts can affect hydropower generation and agricultural electricity use in quite significant ways. However, adding drought indices like the SPEI (Standardised Precipitation-Evapotranspiration Index) would have required collecting and processing an entirely new dataset, which was beyond what the study set out to do. This is something that future research could look into.

Finally, while the study covered a lot of ground by including 47 countries and 22 years of data, the results are still specific to Sub-Saharan Africa. The relationships between electricity demand and its drivers may look quite different in other parts of the world, so the findings cannot simply be copied and applied elsewhere without careful consideration.

1.8 Definition of Key Terms

i. Electricity Demand

The total amount of electrical energy that is needed or consumed in a given area over a certain period of time. In this study, electricity demand is measured in terawatt-hours

(TWh).

ii. **Feature Engineering**

The process of taking raw data and turning it into more useful variables that help a machine learning model make better predictions [12].

iii. **Machine Learning Model**

A type of computer model that is trained using data to find patterns and then uses those patterns to make predictions, without being given explicit rules for every situation [12].

iv. **Econometric Models**

Statistical models that are built on economic theory and mathematics to study the relationships between different variables and to make predictions about outcomes [18].

v. **Climate Variables**

Measurements of environmental conditions, such as average temperature and rainfall, that are associated with how much electricity people use.

vi. **Socioeconomic Indicators**

Variables that describe the economic and social situation in a country or region, such as population size, GDP per capita, urbanisation rate, and the percentage of people with access to electricity.

vii. **Predictive Analytics**

The use of statistical and machine learning methods to analyze data and predict future outcomes.

viii. **Time-Series Data**

Data collected sequentially over time to show how variables change across different time periods.

ix. **Panel Time-Series Data**

A type of dataset that follows the same group of entities, in this case countries, over multiple time periods. This allows researchers to look at both how things change over time and how countries differ from each other.

x. **Forecasting**

The process of predicting future values using historical data and related variables.

xi. **Lag Variables**

New variables created by shifting a variable back in time, so that the model can learn from what happened in previous years. For example, last year's electricity demand can be used as a predictor of this year's demand.

xii. **Multivariate Forecasting**

A way of making predictions that uses several different input variables at the same time, rather than only using the past values of the thing being predicted.

xiii. **RFECV (Recursive Feature Elimination with Cross-Validation)**

A method for selecting the most useful features for a model. It works by removing the weakest features one at a time and uses cross-validation to figure out which number of features gives the best results [19].

xiv. **Mutual Information**

A way of measuring how much knowing the value of one variable tells you about the value of another. Unlike correlation, it can pick up on both straight-line and curved relationships between variables.

xv. **SHAP (SHapley Additive Explanations)**

A method for explaining what a machine learning model is doing. It uses ideas from game theory to show how much each input variable contributed to a particular prediction [20].

xvi. **Walk-Forward Validation**

A way of testing a time-series model that respects the order of time. The model is trained on past data, makes a prediction for the next period, and then that period is added to the training data before the next prediction is made [13].

xvii. **Permutation Importance**

A technique used after a model has been trained to measure how important each feature actually is. It works by randomly scrambling the values of one feature at a time and seeing how much the model's performance drops as a result.

Chapter 2

Literature Review

2.1 Introduction

This chapter reviews what other researchers have done in the area of electricity demand forecasting, with a particular focus on machine learning methods, the use of climate and socioeconomic variables, and studies that are relevant to Sub-Saharan Africa. The chapter also looks at the challenges that come with this kind of research and identifies the gaps that this study was designed to address.

Electricity demand forecasting has received a lot of attention in recent years, mainly because energy systems around the world are becoming more complex and the demand for electricity keeps growing [1]. Machine learning methods have been increasingly adopted in this area because they are able to model the kinds of complicated, nonlinear relationships that exist between electricity demand and the many factors that influence it, things like climate conditions, population growth, and economic activity [4, 12, 11]. At the same time, the quality of the forecasts still depends heavily on how good the data is, how well different datasets are combined, and how carefully the input features are chosen and built.

This chapter is organised as follows. It starts by explaining the different categories of electricity demand forecasting. It then reviews the main forecasting approaches that have been used, including traditional statistical methods and machine learning methods. After that, it looks at how climate and socioeconomic variables have been incorporated into forecasting models. The challenges that researchers have faced are then discussed, followed by a comparison of key studies from the literature and a critical analysis of their strengths and weaknesses. The chapter ends by identifying the research gaps that motivated this study.

2.2 Types of Electricity Demand Forecasting

Electricity demand forecasting is generally grouped into three broad categories based on how far ahead the forecast is looking: short-term, medium-term, and long-term [4].

Short-term forecasting covers a period ranging from a few hours to a few days ahead. This kind of forecasting is used mainly for managing the electricity grid in real time, making sure that the right amount of power is being generated at any given moment, and responding quickly to sudden changes in demand [7]. Because it deals with very recent data and very short time horizons, short-term forecasting often involves hourly or sub-hourly data and relies heavily on weather forecasts and real-time electricity consumption data [21].

Medium-term forecasting typically looks weeks to months ahead. It is used for things like scheduling maintenance on power plants, planning fuel purchases, and distributing resources across different parts of the energy system more efficiently. This kind of forecasting helps energy providers avoid shortages or surpluses in the near future without having to make long-term commitments [4].

Long-term forecasting looks several years or even decades into the future. This is the most relevant category for energy planning and policy making, particularly in developing regions. Long-term forecasts inform decisions about where to build new power plants, how to expand the electricity grid, which energy sources to invest in, and how to plan for population growth and economic development [7, 8]. In Sub-Saharan Africa, where energy infrastructure is still being built out and where the energy needs of the population are expected to grow substantially over the coming decades, long-term forecasting is especially important [2].

This study falls into the long-term forecasting category. It used annual country-level data and focused on understanding and predicting how electricity demand changes over multiple years, across a wide range of countries in Sub-Saharan Africa.

2.3 Electricity Demand Forecasting Approaches

2.3.1 Traditional Statistical Methods

For a long time, the main tools used for electricity demand forecasting were traditional statistical and econometric methods. These included things like autoregressive integrated moving average (ARIMA) models, linear regression, and various forms of time series analysis [4]. These methods worked reasonably well in situations where the data behaved in a predictable, linear way and where there were not too many variables to deal with at once.

However, as the factors influencing electricity demand became more numerous and more complicated, the limitations of these approaches started to show. Linear models, by definition, cannot capture relationships that are not straight lines. They also tend to struggle when variables interact with each other in complex ways, or when the patterns in the data change over time [4]. For electricity demand in a region like Sub-Saharan Africa, where demand is being driven by a rapidly changing mix of population growth, economic development, and shifting climate patterns, these limitations are particularly significant.

Econometric models, including panel data regression models like Fixed Effects and Random Effects models, have also been widely used in energy research [18]. These models have the advantage of being able to account for differences between countries or regions that do not change over time, such as geography or institutional factors. However, they still rely on linear relationships and do not deal well with complex nonlinear dynamics [22].

2.3.2 Machine Learning Approaches

Machine learning methods have become increasingly common in electricity demand forecasting because they are able to pick up patterns in data that traditional statistical methods cannot [12]. Unlike regression-based models, machine learning algorithms do not assume that the relationship between variables follows a specific mathematical form. Instead, they learn the patterns directly from the data, which means they can handle both linear and nonlinear relationships, as well as interactions between variables [11].

Some of the most commonly used machine learning models in electricity demand forecasting

include Random Forest, Gradient Boosting, XGBoost, Support Vector Machines (SVM), Artificial Neural Networks (ANN), and various ensemble and deep learning methods [10, 23]. Each of these has its own strengths and is better suited to some types of problems than others.

Random Forest builds a large number of decision trees and combines their predictions, which makes it less sensitive to noise and less likely to overfit [24]. This makes it a good choice for situations where the dataset has a lot of variables and where some of those variables might be correlated with each other. Gradient Boosting and XGBoost work in a similar way but build their trees one at a time, with each new tree trying to correct the mistakes of the previous ones [25]. XGBoost in particular has become very popular in the forecasting literature because it tends to be both accurate and computationally efficient [26]. LightGBM is another gradient boosting method that is designed to be even faster than XGBoost, especially on large datasets, by using a leaf-wise rather than a depth-wise tree growing strategy [15].

Deep learning models, including Long Short-Term Memory networks (LSTM) and convolutional neural networks, have also been applied to electricity demand forecasting with some success, particularly for short-term forecasting where large amounts of high-frequency data are available [27, 28]. However, these methods tend to require much more data than is typically available at the country level and are generally harder to interpret, which makes them less suitable for the kind of long-term, multi-country analysis carried out in this study.

2.4 Integration of Climate and Socioeconomic Variables

One of the things that has become clear from the research literature is that electricity demand is not just determined by its own past values. It is also shaped by a wide range of external factors, including climate conditions and the economic and social situation of the country or region in question [11, 2].

2.4.1 Climate Variables

Temperature is probably the most widely used climate variable in electricity demand forecasting. The reason is straightforward: when temperatures are very high, people use more electricity for air conditioning and cooling, and when temperatures are very low, they use more for heating [10]. A number of studies have confirmed this relationship across different regions and datasets [29, 30]. In Sub-Saharan Africa, however, the relationship between temperature and electricity demand is somewhat different from what is seen in temperate regions. Because much of the region is already warm, and because many households and businesses do not have air conditioning, the cooling effect of temperature on demand is less pronounced, and high temperatures may sometimes even be associated with lower demand in countries where electrification is limited.

Precipitation is another climate variable that has been included in some forecasting studies. Rainfall can affect electricity demand in several ways. In countries that rely heavily on hydropower, rainfall affects how much electricity can be generated, which in turn affects supply and demand dynamics. Rainfall also influences agricultural activity, which is a significant consumer of electricity in some parts of Sub-Saharan Africa [1].

More recent work has also looked at derived climate variables, such as temperature anomalies (how much a given year's temperature differs from the long-run average), cooling degree days (a measure of how much cooling energy is needed), and rolling averages of climate indicators [31]. These kinds of features can capture things that raw temperature or rainfall values on their own cannot, such as unusually hot years or persistent droughts.

2.4.2 Socioeconomic Variables

Economic and social conditions are also major drivers of electricity demand. A growing economy means more industrial and commercial activity, which requires more electricity. A larger population means more households consuming electricity for lighting, cooking, and appliances. Urbanisation concentrates demand in cities, where electricity infrastructure is better developed and consumption per person tends to be higher. And as more people gain access to electricity for the first time, overall demand rises even if individual consumption stays the same [2, 11].

Several studies have used GDP per capita as a proxy for economic development and found a strong positive relationship with electricity demand [22, 23]. Population size and urbanisation rate have also been widely used predictors in long-term electricity demand models [10, 29]. Electricity access rate, which measures the share of the population that has a connection to the electricity grid, is particularly relevant in Sub-Saharan Africa, where large parts of the population are still being connected to the grid for the first time [2].

What is less common in the literature is the use of growth rate versions of these variables, which capture how fast things are changing rather than just how big they are. This distinction is important in a region like Sub-Saharan Africa where the rate of change in population, urbanisation, and electricity access can be just as important as the current level of these indicators.

2.5 Challenges in Electricity Demand Forecasting

Several recurring challenges appear across the electricity demand forecasting literature, and these were all relevant to this study as well.

2.5.1 Data Quality and Integration

One of the biggest practical challenges in electricity demand forecasting is the quality and availability of data. Electricity data, climate data, and socioeconomic data are all collected by different organisations, stored in different formats, and cover different time periods and geographies [9]. Putting them all together into a consistent, analysis-ready dataset requires a significant amount of work, and there are many opportunities for errors or inconsistencies to creep in [17].

This problem is particularly acute in Sub-Saharan Africa, where data collection systems are less well developed than in higher-income countries. Some countries have missing data for certain years or certain variables, and the way data is defined and measured can vary from one country to another. This makes it hard to do the kind of multi-country analysis that is needed to understand electricity demand patterns across the region as a whole [8].

2.5.2 Model Overfitting and Generalisation

Another common challenge is the risk of overfitting, which is when a model learns the patterns in the training data so well that it cannot generalise to new, unseen data. This is a particular concern with complex machine learning models that have many parameters and can fit almost any pattern in the training data [12]. In the context of electricity demand forecasting, overfitting can lead to models that appear to perform well during training but then fail when used to make actual forecasts.

Several techniques have been used to address this problem, including cross-validation, regularisation, and walk-forward validation [13, 32]. Walk-forward validation is particularly well suited to time series problems because it respects the chronological order of the data and prevents future information from leaking into the training process.

2.5.3 Feature Selection

Choosing which variables to include in a forecasting model is also a non-trivial challenge. Including too many variables can lead to overfitting and can make the model harder to interpret. Including too few means that important information is being left out. The machine learning literature has developed a range of techniques for selecting the most useful features, including correlation analysis, recursive feature elimination, and tree-based importance measures [19, 24].

In the context of electricity demand forecasting, an additional complication is that the most informative features for long-term forecasting may be different from those that matter for short-term forecasting. Lag features and growth rate features, for example, are more relevant for capturing multi-year trends than for predicting next week’s demand [12].

2.5.4 Geographic Focus of Existing Research

Perhaps the biggest gap in the existing literature, from the perspective of this study, is the heavy geographic concentration of electricity demand forecasting research. The vast majority of studies have been conducted in high-income countries or in countries like China and India where high-quality electricity data is widely available [7, 8]. Research focused specifically on Sub-Saharan Africa is much rarer, and the studies that do exist tend to look at a single country rather than the region as a whole.

2.6 Comparative Analysis of Existing Studies

Table 2.1 summarises a selection of key studies from the electricity demand forecasting literature, along with the methodology used, the data source, the reported performance, and the main limitations of each study.

Table 2.1: Comparative Analysis of Related Studies

Study	Methodology	Dataset	Performance	Limitations
Tang et al. (2021)	Regression-based ML models including gradient boosting	Household energy consumption data	ML models outperformed statistical baselines, with MAPE ranging from 3–12% depending on forecasting horizon and selected features.	Focused on building-level short-term demand and is not suitable for country-level or multi-country forecasting.
Ahmad et al. (2021)	Systematic review of RF, SVM and ANN forecasting models	Multiple datasets from different regions	Reviewed studies reported MAPE values between 2–15%, with ensemble models generally outperforming linear and statistical methods.	Review paper only; findings are aggregated across many studies rather than a single dataset or application.
Shah et al. (2022)	Random Forest and Gradient Boosting using climate variables	Electricity consumption and weather data	Random Forest achieved MAPE below 5% for short-term forecasting when climate variables were included.	Limited to short-term forecasting within a single country; results may not generalise to long-term panel forecasting.
Hong & Fan (2016)	Probabilistic load forecasting using quantile regression and simulation	Historical electricity demand datasets	Probabilistic methods reduced prediction interval width by 10–30% compared with point forecasting.	Focused on uncertainty estimation rather than incorporating climate or socioeconomic drivers.
Wang et al. (2022)	Ensemble learning models	Short-term electricity and climate data	Ensemble models reduced RMSE by 8–12% and achieved R^2 above 0.94 across evaluated datasets.	Only evaluated short-term forecasting; long-term panel forecasting was not investigated.
Rahman et al. (2022)	Comparison of RF, XGBoost and SVM	Multiple energy demand datasets	Random Forest achieved R^2 above 0.90 in most datasets and consistently outperformed ARIMA and linear regression.	Most studies relied mainly on historical demand and weather data with limited socioeconomic variables.
Török et al. (2025)	Multiple Linear Regression and Fixed Effects Panel Model	Eurostat panel data (27 EU countries, 2015–2023)	Fixed effects explained 78% of variation in renewable energy share ($R^2=0.78$); investment rate remained statistically significant.	Examined renewable energy share rather than electricity demand and focused on the European Union.
World Bank (2023)	Descriptive analysis of energy access and consumption	Global socioeconomic and energy access data	Reported that over 600 million people in Sub-Saharan Africa lack electricity access and businesses lose about 8% of revenue because of power disruptions.	Descriptive study only; no predictive or forecasting models were trained.

2.7 Critical Analysis of Existing Studies

Looking across the studies reviewed, a few consistent findings stand out. First, machine learning methods, particularly ensemble methods like Random Forest and XGBoost, generally outperformed traditional statistical and econometric methods when applied to electricity demand forecasting [12, 10]. This was true across a range of different datasets and contexts, which suggests that the ability of machine learning to capture nonlinear patterns in the data is genuinely valuable in this domain.

Second, including climate and socioeconomic variables alongside historical electricity demand

data tended to improve forecast accuracy, though the size of the improvement varied depending on the study and the context [10, 23]. In studies where the relationship between electricity demand and these external variables was strong and consistent, the improvement was substantial. In other cases, particularly when the dataset was small or noisy, the improvement was more modest.

However, there were also some consistent weaknesses in the existing literature. The majority of studies focused on short-term forecasting, using hourly or daily data, and the findings from these studies do not necessarily transfer to the long-term, annual forecasting setting that is most relevant for energy planning in Sub-Saharan Africa [7]. Most studies were also conducted in a single country, which limits how broadly the findings can be applied.

Feature engineering was another area where existing studies often fell short. Many studies used only the raw values of the predictor variables and did not explore whether creating derived features, like lag variables, growth rates, or anomaly measures, could improve the models [23, 26]. Given how important these kinds of temporal features are for capturing multi-year trends in electricity demand, this is a notable gap.

Finally, very few studies made any attempt to explain their models or help decision-makers understand what is driving the forecasts. Most papers reported accuracy metrics and stopped there, without providing the kind of feature importance analysis or visual decision-support tools that would make the results more useful to energy planners and policymakers [33, 34].

2.8 Research Gap

Based on the review of the existing literature, several clear research gaps were identified, all of which this study was designed to address.

The first gap was the lack of integrated multi-country forecasting models for Sub-Saharan Africa. Most studies that included climate and socioeconomic variables were either single-country studies or focused on regions outside Sub-Saharan Africa [7, 8]. This meant that very little was known about how electricity demand patterns differ across countries in the region, or about which factors are most consistently important across a wide range of contexts. Deb et al. [7] produced strong short-term results but worked within a single national grid, so nothing in their findings tells a planner whether the same drivers hold one border away. Ali et al. [8] faced the same limit. This study answers that directly by fitting one model across 47 countries and reporting which drivers hold across all of them.

The second gap was the absence of comprehensive, multi-source integrated datasets for electricity demand forecasting in Sub-Saharan Africa. Electricity data, climate data, and socioeconomic data were available from various sources, but no study had brought them all together into a single, consistent, country-year panel dataset that covered the region over a long time period [10, 31]. Shah et al. [10] showed the value of combining climate with demand data, but assembled that combination for one setting only. No equivalent country-year panel existed for Sub-Saharan Africa, which is why building the integrated dataset became an output of this study in its own right and not just a preparatory step.

The third gap was in feature engineering. Most existing studies did not go beyond using raw variable values, and very few explored the use of lag features, growth rate features, interaction features, or climate anomaly features in the context of long-term electricity demand forecasting [23, 26]. This study was therefore designed to explore how a wider range of engineered features could improve forecasting interpretability and support structural analysis of demand patterns. Rahman et al. [23] and Singh et al. [26] both fed raw variable values into their models. Neither

tested whether a rolling average or a growth rate would carry more signal than the level. This study tests exactly that, and the finding that the three-year rolling mean of population outranks raw population on both tree-based importance measures, while ranking below it on the correlation and RFECV measures, shows the question was worth asking.

The fourth gap was in model comparison. While many individual studies evaluated one or a small number of machine learning models, few studies compared multiple model types, including both traditional statistical models and machine learning models, using the same dataset and the same validation procedure [12, 23]. This made it difficult to draw general conclusions about which approaches work best for electricity demand forecasting in Sub-Saharan Africa. Ahmad et al. [12] compared machine learning models against each other but did not include panel econometric models, which are the natural choice for country-year data. This study places both families on the same data with the same validation, which is what makes the comparison meaningful.

The fifth gap was in model explainability and decision support. Few studies combined feature importance analysis with an interactive dashboard to help policymakers, researchers, and energy planners understand both the forecast results and the factors that were driving them [33, 34]. This study addressed this gap by developing a Streamlit dashboard that presented forecasting results, demand trends, and feature importance in an accessible, interactive format. Zhang et al. [33] noted that accuracy metrics on their own do not help a planner decide anything. Very few studies act on that observation. This study responds by pairing permutation importance with a dashboard, so that the drivers behind a forecast are visible alongside the forecast itself.

2.9 Chapter Summary

This chapter reviewed the electricity demand forecasting literature, covering the main types of forecasting, the methods that have been used, the role of climate and socioeconomic variables, and the challenges that researchers have faced. The review showed that machine learning methods, particularly ensemble approaches like Random Forest and XGBoost, generally performed better than traditional statistical methods, and that including climate and socioeconomic variables alongside historical demand data was associated with better forecast accuracy in some contexts, though the size of the improvement varied considerably.

At the same time, the review identified several important gaps in the existing literature: the lack of multi-country studies in Sub-Saharan Africa, the absence of integrated multi-source datasets, limited attention to feature engineering, limited model comparison using consistent datasets and validation procedures, and limited effort to make forecasting results accessible and useful to decision-makers. It should also be noted that machine learning is a rapidly evolving field, and some methods and findings that were recent at the time of writing may be superseded by newer approaches. The sources cited in this review were selected to be as current as possible, but future work should continue to incorporate new literature as it emerges. These five gaps shaped the study directly. The first two set the geographic and data scope, the third set the feature engineering work, the fourth set the six-model comparison, and the fifth set the dashboard. Chapter 3 describes how each was carried out.

Chapter 3

Methodology

3.1 Introduction

This chapter describes the methods that were used to achieve the objectives of the study. The study tested whether climate and socioeconomic data from Sub-Saharan African countries improves annual electricity demand forecasts, using six modelling approaches and three feature sets. Accurate electricity demand forecasting is important because it supports energy planning, resource allocation, and infrastructure development [1, 4].

The chapter is organised to follow the order in which the work was carried out. It begins with the research approach and design, then covers the data sources, integration and preprocessing, feature engineering and selection, model training and evaluation, and the steps taken to prevent data leakage. Section 3.3 sets out the full analytical pipeline.

3.2 Research Approach

This study followed a quantitative approach. All the evidence used was numerical, drawn from secondary sources, and all conclusions rested on measured relationships between variables rather than on interpretation of accounts or opinions. The reasoning was deductive: the study began with claims from the existing literature about which climate and socioeconomic factors drove electricity demand, turned those claims into measurable features, and then tested them against observed demand.

Within the quantitative family, the approach was predictive rather than explanatory in the causal sense. The study did not claim to establish that a change in population caused a change in demand. It measured how strongly each variable was associated with demand and how much each one improved a forecast. This distinction mattered for how the findings should be read, and it was returned to in Chapter 5.

A quantitative approach was the right choice here because the research questions were about magnitude and comparison. Asking how much accuracy improves when climate variables are added requires measurement on a common scale across all 47 countries. A qualitative approach could describe how planners made forecasting decisions, but it could not answer these particular questions.

3.3 Research Design

The study used a longitudinal panel design with a comparative modelling structure.

The design was longitudinal and panel-based because the unit of analysis was the country-year. Each of the 47 countries contributed one observation per year from 2000 to 2021, giving repeated measures on the same units over time. The modelling frame used a shorter window, 2002 to 2021, for the reason given in Section 3.11. This structure allowed the study to look at how demand changed within a country over the years and how demand differed between countries, and it was the reason panel regression models were included alongside the machine learning models.

The design was ex-post facto. All variables were observed as they occurred, and nothing was manipulated. There was no treatment group and no control group, because the study was not testing an intervention. It was testing how well a set of already-recorded variables predicted an already-recorded outcome.

The comparative element was the core of the design. Rather than building one model and reporting its accuracy, the study defined three feature sets and ran the same six modelling approaches and the same validation procedure across all of them. Model A used past demand only. Model B used climate and socioeconomic variables only. Model C used both. Because everything except the feature set was held constant, the difference in performance between Model A and Model C isolated exactly what the climate and socioeconomic variables contributed, and Model B showed what they could do on their own. This three-way structure was what allowed the study to answer Research Question 2 rather than simply report an accuracy figure.

Time was handled by a chronological split rather than a random one. The models learned from 2002 to 2017 and were tested on 2018 to 2021, so that the test always sat in the future relative to the training data. This mirrored the real forecasting situation and was described in detail in Section 3.11.

3.3.1 Forecasting Strategy

A one-step-ahead forecasting approach was used, where electricity demand for the following year was predicted using historical observations from previous years. A direct forecasting strategy was applied, meaning that the models predicted future electricity demand directly from the available input variables rather than using previously predicted values as inputs.

This approach was appropriate for the study because the data was annual and the goal was to understand how structural factors like population, GDP, and climate conditions drive demand year by year. It also helped to avoid the error accumulation problem that can occur in recursive forecasting approaches, where errors from one prediction are carried forward into the next [7].

3.3.2 Analytical Framework

The study followed a structured analytical pipeline with seven main phases. First, data on electricity demand, climate, and socioeconomic factors was gathered from trusted public sources. The collected data was then cleaned, aligned, and merged into a single panel dataset using country-year identifiers. After preprocessing, feature engineering techniques were applied to create a rich set of predictors. Feature selection methods were then used to identify the most informative features. Machine learning forecasting models were trained and evaluated using standard performance metrics. Walk-forward validation was applied to ensure the models

generalised well to unseen data. Finally, the forecasting results were presented through an interactive Streamlit dashboard [10].

3.4 Data Sources and Dataset Description

3.4.1 Data Sources

The study used publicly available datasets collected from trusted international sources. Electricity demand data was obtained from the World Energy Consumption dataset published by Our World in Data. Socioeconomic indicators including GDP per capita, population, urbanisation rate, electricity access rate, and urban electricity access rate were collected from the World Bank Open Data portal [2].

Climate variables were obtained from the ERA5 reanalysis dataset, accessed through the World Bank Climate Change Knowledge Portal (CCKP). ERA5 is produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) and provides high-quality, globally consistent climate records at a spatial resolution of 0.25 degrees [1]. ERA5 was selected because it offers better temporal coverage for the full 2000 to 2021 study period, is available in a pre-aggregated annual country-level format through the CCKP portal, and is widely used and cited in the energy and climate research literature.

3.4.2 Dataset Description

The initial integration produced a dataset of 1,056 rows covering 48 Sub-Saharan African countries. South Sudan was deliberately kept in the dataset at this stage rather than being dropped right away. The reason was to allow its data coverage and missing value patterns to be compared against the other 47 countries during the exploratory analysis. This comparison confirmed that South Sudan had more than a decade of completely missing data (2000 to 2010) across all variables, since the country only became independent in 2011. Unlike other countries that had a few scattered missing values that could be filled in using interpolation, South Sudan's gaps were too large and too concentrated at the start of the study period for any imputation method to produce reliable values. Filling in eleven consecutive years of data for every variable using only a handful of post-2011 observations would have introduced far more noise than signal into the dataset. South Sudan was therefore removed after the exploratory analysis, reducing the dataset to 1,034 rows across 47 countries. The descriptive statistics in Table 4.1 were computed before this exclusion, which is why some variable counts in that table are slightly above 1,034.

The dataset followed a panel time-series format, meaning that the same group of countries was observed repeatedly across all 22 years of the study period. This structure allowed the study to capture both how individual countries changed over time and how countries compared with each other at any given point in time. All datasets were merged using country ISO code and year as joint identifiers to ensure that every observation was correctly matched across sources.

3.4.3 Variable Selection and Justification

The target variable for this study was electricity demand, measured in terawatt-hours (TWh). This variable was chosen because it directly represents the amount of electricity consumed within a country's energy system in a given year, and it is the quantity that energy planners and policymakers need to forecast in order to make sound infrastructure and resource allocation decisions.

Climate variables included average annual temperature and total annual precipitation (rainfall). Temperature was selected because research consistently shows that it influences electricity usage through heating and cooling needs across different seasons and climates [10]. Precipitation was selected because it affects seasonal agricultural activity, which is a significant electricity consumer in many Sub-Saharan African countries, and because it influences the output of hydropower systems that supply electricity to a large proportion of the region [1].

Socioeconomic variables included GDP per capita, total population, urbanisation rate, electricity access rate (the percentage of the total population with access to electricity), and urban electricity access rate (the percentage of the urban population with access to electricity). GDP per capita was selected because economic activity is one of the strongest long-term predictors of electricity demand, as growing economies need more energy for industry, commerce, and services [2, 11]. Population was selected because the total number of people in a country directly influences how much electricity is consumed in homes and businesses. Urbanisation rate was included because urban areas tend to consume significantly more electricity per person than rural areas, and because rapid urbanisation in Sub-Saharan Africa is one of the key structural correlates of demand growth [2, 11]. Electricity access indicators were included because the process of connecting previously unelectrified households and businesses directly adds to total demand, and because this process is still very much ongoing in much of the region [2].

Variables such as electricity prices, government energy policies, political stability, and consumer behaviour were considered but ultimately excluded from the study. While these factors are known to influence electricity demand, consistent data for all 47 countries across the full 22-year study period was not available through publicly accessible sources. The variables that were selected were specifically chosen because they met two criteria: they are theoretically important drivers of electricity demand based on the existing literature, and they are consistently available across all countries and years in the study [10, 11].

Table 3.1 summarises the variables used in the study, their categories, data sources, and the justification for their inclusion.

Table 3.1: Study Variables and Justification

Variable	Category	Source	Justification
Electricity Demand (TWh)	Target Variable	Our World in Data (World Energy Consumption)	Represents total electricity consumption; the quantity to be forecast for energy planning purposes
Average Temperature (°C)	Climate Variable	ERA5 / World Bank CCKP	Influences electricity usage through heating and cooling needs across seasons
Precipitation (mm)	Climate Variable	ERA5 / World Bank CCKP	Influences seasonal agricultural electricity use and hydropower output
GDP per Capita (USD)	Socioeconomic Variable	World Bank	Represents economic activity and the capacity for industrial, commercial, and household electricity consumption
Population	Socioeconomic Variable	World Bank	Directly determines the scale of household and commercial electricity demand
Urbanisation Rate (%)	Socioeconomic Variable	World Bank	Captures urban growth and associated expansion in electricity infrastructure and per-capita consumption
Electricity Access Rate (%)	Socioeconomic Variable	World Bank	Measures the share of the population connected to the electricity grid; expanding access directly increases total demand
Urban Electricity Access Rate (%)	Socioeconomic Variable	World Bank	Captures electricity access specifically in urban areas, where consumption tends to be higher

3.5 Data Integration and Preprocessing

3.5.1 Data Integration

The seven datasets collected from different sources were integrated into a single panel dataset using country ISO code and year as joint identifiers. A complete country-year grid was first constructed for all 47 countries and all 22 years, and then each dataset was merged onto this grid using left joins. This approach ensured that every country-year combination was represented in the final dataset, even in cases where some variables had missing values for a particular country or year.

Climate data from the ERA5 dataset was downloaded in annual format directly from the World Bank CCKP portal, meaning no monthly aggregation was needed. The ERA5 data used the convention of anchoring annual values to July of each year (e.g., the column ‘2005-07’ represented the annual average for 2005), which was accounted for during the merging

process.

3.5.2 Outlier Treatment

Outlier detection was performed using the interquartile range (IQR) method, which identifies values that fall beyond 1.5 times the IQR below the first quartile or above the third quartile. Boxplots were produced for all variables to visualise the distribution of outlier values. However, upon closer inspection of the individual outlier values, these values were found to represent genuine country-level differences rather than data or measurements errors. For example, South Africa's electricity demand of over 200 TWh, Nigeria's population of over 200 million, and Mauritius's GDP per capita of over USD 10,000 are all real values that reflect the true economic and demographic diversity of the region. Removing or capping these values would have distorted the dataset and hidden real patterns that the models needed to learn from. Therefore, no outlier removal or capping was applied.

3.5.3 Normality Testing

Before deciding which variables needed transformation, the Shapiro-Wilk test [35] was used to test each numeric variable for normality. The Shapiro-Wilk test was selected because it performs well for small to moderate sample sizes and is widely used in applied forecasting research. All variables failed the normality test at the 5% significance level, confirming that none of the variables followed a normal distribution. Skewness values were also computed for each variable. Three variables showed high right-skewness with absolute skewness values greater than 1: electricity demand (skewness = 6.44), population (skewness = 3.54), and GDP per capita (skewness = 2.95). These three variables were flagged as candidates for log transformation, which was applied during the feature engineering stage.

3.5.4 Missing Value Treatment

After integration, the dataset had a small number of missing values. The worst affected were GDP per capita with 24 missing values, electricity access rate with 16, electricity demand with 12, and urban electricity access rate with 7. Together these came to less than 3% of any single variable. These gaps were handled using linear interpolation within each country's own time series. Linear interpolation was chosen because it fills missing values by drawing a straight line between the known values on either side of the gap, which preserves the trend of each country's own data without borrowing information from other countries [17]. For gaps at the very start or end of a country's time series, where interpolation could not work because there was no value on one side of the gap, forward fill was applied to carry the last known value forward in time. Backward fill was deliberately avoided. Pulling a later value backward into an earlier gap could move information from the test period (2018 to 2021) into earlier training years, which would be a form of data leakage. Using forward fill only makes sure that gaps are filled with past information alone.

3.5.5 Scaling

Z-score standardisation was applied to the predictor variables to bring them all onto the same scale. This type of scaling works by taking each value, subtracting the mean of that variable, and then dividing by the standard deviation [16]. The formula is:

$$z = \frac{x - \mu}{\sigma}$$

where x is the original value, μ is the mean, and σ is the standard deviation. After scaling, each variable ends up with a mean of zero and a standard deviation of one, regardless of what units it was originally measured in.

This step was mainly needed for the Linear Regression model, which is sensitive to differences in the scale of input variables. If one variable is measured in millions (like population) and another is measured in percentages (like electricity access), Linear Regression would give more weight to the one with bigger numbers, not because it matters more but simply because its values are larger. Scaling removes this problem by putting everything on the same footing. Tree-based models like Random Forest and XGBoost are not affected by scale because they split on thresholds rather than magnitudes, but scaling was applied to all variables for consistency across the pipeline [13].

One important note is how the mean and standard deviation were calculated. If they were calculated using the entire dataset including the test period (2018 to 2021), then information from the future would leak into the training data through the scaling. To prevent this, the scaler was fitted only on the training period data (2000 to 2017). This means the mean and standard deviation used for the scaling came entirely from training-period values. These same training-period parameters were then applied to scale the test-period data as well. So the test rows were scaled using the training mean and training standard deviation, not their own.

This was all done using scikit-learn's `StandardScaler`. The `fit()` method was called on the training-period rows only, which calculated and stored the mean and standard deviation for each variable. Then the `transform()` method was called on the full dataset (including test rows), which applied the stored training-period parameters to every row [36].

3.6 Feature Engineering

Feature engineering was a key stage of the methodology. Instead of relying only on the original variables, a broader set of input features was created from the available data. This was necessary because raw annual values of variables such as population and GDP per capita do not fully capture how electricity demand changes over time. Features that represent rates of change, recent trends, deviations from normal conditions, and interactions between variables can provide additional structural information that may support model interpretability and analysis [12, 10].

The predictor variables were grouped into two main categories: climate variables and socioeconomic variables. Beyond these original variables, additional predictors were derived through feature engineering by applying transformations such as logarithmic scaling and rolling averages, creating interaction features between related variables, and calculating year-on-year growth rates.

Overall, 27 engineered features were created and organised into nine feature engineering categories, as described in the following subsections. The log-transformed target variable is counted separately and is not included in this total, since it is the quantity being predicted rather than a predictor.

3.6.1 Log Transformation

Log transformation was applied to electricity demand, population, and GDP per capita, which were the three variables identified as highly positively (right) skewed by the normality tests, with skewness values of 6.44, 3.54, and 2.95 respectively, all well above the common threshold of

1 that indicates high positive skew. The specific transformation used was $\log(1+x)$, also written as $\log_{1p}(x)$, rather than the standard natural logarithm. This form was chosen because it is safe for values close to zero, which was relevant for electricity demand given that some smaller countries had very low TWh values. The transformation reduced the skewness of electricity demand from 6.44 to 1.56, population from 3.54 to -0.31, and GDP per capita from 2.95 to 0.53, making all three distributions substantially more symmetric and better suited for linear modelling.

The log-transformed version of electricity demand (`log_electricity_demand`) was used as the target variable throughout the modelling process, and all model evaluation metrics were calculated on this same log scale. This kept the comparison between models and feature sets consistent, since every model was predicting the same quantity. It does however mean that the MAE and RMSE values reported in Chapter 4 are in log units rather than terawatt-hours, and that the MAPE values are percentages of the log value rather than percentages of actual demand. The inverse transformation $\text{expm1}(x) = e^x - 1$ was applied only inside the dashboard, so that the demand values displayed to users were shown in terawatt-hours.

3.6.2 Lag Features

Lag features were created from the log-transformed electricity demand variable. A one-year lag (`lag_1`) and a two-year lag (`lag_2`) were created within each country's own time series, meaning that no country borrowed lagged values from another. These features were included because electricity demand exhibits strong temporal persistence, meaning that demand in a given year is heavily influenced by demand in previous years [13, 31]. The one-year lag was also required as an explanatory variable in the Dynamic Fixed Effects Panel Regression model. The lag features were computed after log transformation, so they represent lagged log-demand values.

3.6.3 Growth Rate Features

Year-on-year percentage growth rate features were created for five variables: GDP per capita, population, electricity access rate, electricity demand, and urbanisation rate. The growth rate for each variable was computed as:

$$\text{growth rate}_t = \frac{x_t - x_{t-1}}{x_{t-1}} \times 100$$

where x_t is the value in year t and x_{t-1} is the value in the previous year [31]. These features captured how fast key indicators were changing from one year to the next, rather than just their absolute size. The inclusion of electricity demand growth rate was justified because demand growth momentum can itself be informative about future demand. Urbanisation growth rate was included because the pace of urbanisation, not just its level, directly affects how quickly electricity infrastructure needs to expand. GDP and population growth rates were included for similar reasons [23, 26].

3.6.4 Interaction Features

Three interaction features were created to represent the combined effect of two variables on electricity demand, where the effect of one variable was expected to depend on the level of another:

- **temp_x_urbanisation:** The product of standardised average temperature and standardised urbanisation rate. This feature was included because hot temperatures in highly

urbanised areas are likely to generate stronger cooling-driven electricity demand than the same temperature in a rural area [10].

- **gdp_x_electricity_access**: The product of standardised log GDP per capita and standardised electricity access rate. Economic growth drives electricity demand primarily in areas where people and businesses actually have access to the grid. In countries with low electricity access, high GDP does not translate as directly into electricity consumption [2].
- **pop_x_electricity_access**: The product of standardised log population and standardised electricity access rate. A large population only increases electricity demand to the extent that those people are connected to the electricity grid.

Interaction features were computed using the standardised (scaled) versions of the base variables to prevent one variable from dominating the interaction term due to its larger numerical scale.

3.6.5 Trend Feature

A linear trend variable was created as:

$$\text{trend} = \text{year} - 2000$$

This variable takes the value 0 in the year 2000, 1 in 2001, and so on up to 21 in 2021. It was included to allow the models to capture the overall upward time trend in electricity demand without using the raw year value, which would have been harder for models to interpret and more prone to extrapolation errors [31].

3.6.6 Climate Anomaly Features

Two climate anomaly features were created to capture how much each year's climate deviated from that country's typical conditions:

- **temp_anomaly**: Each year's average temperature minus the country's long-run average temperature.
- **precip_anomaly**: Each year's total precipitation minus the country's long-run average precipitation.

The long-run average (baseline) was computed using only the training period (2000 to 2017) rather than the full dataset. These features were included because unusually hot or dry years may trigger abnormal changes in electricity demand, particularly for cooling and irrigation, that would not be captured by the raw temperature or precipitation values alone [31].

3.6.7 Rolling Average Features

Three-year rolling average features were created for population, GDP per capita, temperature, and precipitation. The main reason for doing this is that electricity systems don't change overnight. If a country's population suddenly goes up by a million people in one year, that doesn't mean a million new electricity connections appear the next day. It takes time to build power lines, set up substations, and actually connect new homes and businesses. So what is most strongly associated with demand is not what happened in one single year, but the general trend over the past few years. The rolling average captures this by averaging the current

year together with the two years before it, which gives a smoother picture of where things are actually heading [12]. Same thing applies for GDP (economic growth takes a while before it shows up as new factories using electricity), temperature (one weird hot year doesn't change the whole demand pattern), and precipitation (seasonal rain effects tend to even out over a few years).

Each rolling average was calculated within each country separately, using a backward-looking window. That means for any given year, only that year and the two before it were used. Nothing from the future was ever included. For example, using Angola's actual population data from the study:

- **2000:** No earlier data existed, so the value was just the 2000 population on its own: 16,194,870.
- **2001:** Two years available, so it was the average of 2000 and 2001: $(16,194,869 + 16,747,208) / 2 = 16,471,039$.
- **2002:** First proper 3-year average: $(16,194,869 + 16,747,208 + 17,327,699) / 3 = 16,756,592$.
- **2003 and after:** Always used the current year plus the two before it. So 2003 used 2001, 2002, 2003. And 2004 used 2002, 2003, 2004, and so on.

Three years was picked as the window size because it's long enough to smooth out the noise but short enough to still catch real changes. Going longer, like five or ten years, would have over-smoothed things and the model would have missed actual shifts that were happening.

Because the window only ever looks backward, there is no data leakage here. A training row like 2017 only uses 2015, 2016, and 2017 in its calculation, all training years. Nothing from 2018 or later gets involved. Test rows like 2018 do use some training years in their window (2016 and 2017), but that is fine, test rows are allowed to use past information.

3.6.8 Temperature Variability

This feature is about something different from the temperature rolling average. The rolling average shows what the general temperature has been lately. Variability measures how much the temperature has been jumping around. Why does this matter? Because in a country where the temperature stays more or less the same every year, electricity demand is pretty stable and predictable. But in a country where temperatures keep swinging up and down, one year hot then the next year cooler then hot again, the demand for things like cooling and heating keeps changing too, which makes it harder to predict [11]. So this feature helped the model tell apart countries with stable climates from ones with more unpredictable weather patterns.

It was calculated as the three-year rolling standard deviation of temperature, again using a backward-looking window within each country. Using Angola's actual temperatures:

- **2000:** Only one year of data, so no standard deviation could be worked out. You need at least two numbers to measure spread, so this got no value.
- **2001:** Two years available (21.86°C and 21.88°C). Standard deviation was 0.014, meaning temperature barely changed at all.
- **2002:** Three years (21.86, 21.88, 22.07). Standard deviation rose to 0.116 since 2002 was a bit warmer.

- **2005:** Window covered 2003 (22.22°C), 2004 (21.88°C), 2005 (22.43°C). Standard deviation was 0.278, showing temperatures were bouncing around more during this stretch.

Same as the rolling averages, this window only looks backward so there is no leakage. Every training row’s variability value came entirely from training-period temperatures.

3.6.9 Engineered Features Summary

Table 3.2 summarises all 27 engineered features and their categories.

Table 3.2: Summary of Engineered Feature Categories

Category	Engineered Features	Purpose
Log Transformations	log_population, log_gdp_per_capita	Reduce skewness, stabilise variance, and improve model learning.
Lag Features	lag_1_log_electricity_demand, lag_2_log_electricity_demand	Capture temporal dependence by incorporating electricity demand from previous years.
Growth Rate Features	growth_gdp_per_capita, growth_population, growth_electricity_access, growth_electricity_demand, growth_urbanisation_rate	Capture year-on-year changes in socioeconomic and energy indicators.
Interaction Features	temp_x_urbanisation, gdp_x_electricity_access, pop_x_electricity_access	Model combined effects of climate and socioeconomic variables.
Trend Feature	trend	Represent the long-term temporal trend over the study period.
Climate Anomalies	temp_anomaly, precip_anomaly	Measure deviations from each country’s long-term climate averages.
Rolling Statistics	population_rolling_mean_3, gdp_rolling_mean_3, temp_rolling_mean_3, precip_rolling_mean_3, temp_variability	Capture smoothed long-term trends and short-term climate variability using a three-year rolling window.
Scaled Predictors	scaled_avg_temperature, scaled_precipitation, scaled_log_population, scaled_log_gdp_per_capita, scaled_urbanisation_rate, scaled_electricity_access, scaled_electricity_access_urban	Standardise predictor variables using Z-score scaling to ensure comparable feature magnitudes.

3.7 Feature Selection

Feature selection was carried out in two stages. This two-stage approach was necessary because a preliminary finding showed that the lag features completely dominated all selection methods due to their near-perfect correlation with the target variable (Pearson $r = 0.999$ for lag_1 and $r = 0.997$ for lag_2). The selection pool contained 25 of the 27 engineered features. When all 25 features in the selection pool were evaluated together, the lag features received combined scores that were so much higher than everything else that all climate and socioeconomic features were effectively excluded from selection. The two raw log-transformed variables (log population and log GDP per capita) were left out because their standardised versions were used in their place. Since the primary research questions concerned the contribution of climate and socioeconomic variables to electricity demand forecasting, this outcome would have made it impossible to answer those questions. The two-stage approach was therefore adopted to evaluate temporal features and exogenous features separately before combining them.

Table 3.3 traces how the 27 engineered features were reduced to the 7 used in the final model.

Table 3.3: Feature Selection Funnel

Stage	Features In	Features Out	Left	Reason
Engineered features created	27	-	27	Nine feature categories, as shown in Table 3.2.
Remove raw log variables	27	2	25	Raw log population and log GDP per capita were replaced by their standardised versions to avoid duplication.
Split into two groups	25	-	2 lag + 23 exogenous	Lag features had very high correlations with demand ($r = 0.999$ and $r = 0.997$) and dominated the selection methods, so they were assessed separately.
Drop collinear exogenous features	23	4	19	Features with absolute Pearson correlation above 0.75 were identified; the weaker member of each pair was removed.
Rank on five methods and select top five	19	14	5	Pearson, Spearman, mutual information, RFECV, Random Forest and XGBoost scores were normalised and averaged.
Recombine with lag group	5	-	7	Two lag features were combined with five selected exogenous features to form Model C.

The cutoff of five exogenous features was set before the ranking was produced, in order to keep the model simple. There were 747 training observations and six modelling approaches were compared. Using a small number of features helps reduce the risk of overfitting and makes the linear model coefficients easier to interpret. This was important because interpretability was one of the criteria used to select the final model.

The choice of five features was a judgement rather than a threshold obtained from the data. The sixth and seventh ranked features, scaled urbanisation rate (0.420) and urban electricity access (0.399), scored close to the fifth (0.444), so a cutoff of seven would have been equally defensible. Since the three-way comparison later showed that the exogenous features add very little once lagged demand is included, this choice has little bearing on the study’s conclusions. It would matter considerably more for a study focused on Model B alone.

3.7.1 Stage 1: Temporal Feature Evaluation

In the first stage, the two lag features were evaluated against each other using all five selection methods (described below). Both lag features ranked in the top two positions across all methods, confirming that both were consistently important and should be retained. The one-year lag (lag_1) was ranked first by Pearson correlation, RFECV, and Random Forest importance. The two-year lag (lag_2) was ranked first by XGBoost. Both lags were therefore retained as a confirmed temporal feature group.

3.7.2 Stage 2: Exogenous Feature Selection

In the second stage, the lag features were excluded, and the five feature selection methods were applied only to the climate and socioeconomic variables. Before feature selection, a multicollinearity assessment was conducted using Pearson’s correlation coefficient, computed only on the training period (2000 to 2017) to prevent any test-period information from influencing the decision. Feature pairs with an absolute Pearson correlation greater than 0.75 were considered highly collinear. The threshold of 0.75 was chosen because it is a widely used conservative cutoff in the feature selection literature that flags variables sharing more than about 56% of their variance, which is typically considered too high for both to be retained without causing instability in linear models [16]. When a collinear pair was found, the feature with the weaker

correlation with the target variable (log-transformed electricity demand) was removed, keeping the one that provided more predictive value on its own.

Four features were removed due to multicollinearity:

- `scaled_electricity_access` (Pearson $r = 0.838$ with `scaled_electricity_access_urban` and $r = 0.770$ with `scaled_log_gdp_per_capita`). It appeared in two collinear pairs and had a weaker direct correlation with the target than either of its counterparts, so it was removed.
- `gdp_rolling_mean_3` ($r = 0.854$ with `scaled_log_gdp_per_capita`). The rolling mean of GDP and the scaled log GDP per capita carry almost identical information. The scaled log version was kept because it was already available as a standardised predictor and had a slightly stronger direct association with demand.
- `precip_rolling_mean_3` ($r = 0.988$ with `scaled_precipitation`). The rolling mean of precipitation and the annual precipitation value were nearly identical at this time scale, so the raw scaled version was retained as the simpler and more direct variable.
- `scaled_avg_temperature` ($r = 0.998$ with `temp_rolling_mean_3`). Annual temperature and its three-year rolling mean were almost perfectly correlated at country level. The rolling mean was retained because it had the marginally stronger association with the target, and because smoothing across three years reduces the influence of any single unusual year.

After removing these four, 19 exogenous features remained for the selection process.

Note: Random Forest and XGBoost were the only two models used for tree-based feature importance scoring. This is because they are the only models in the study that naturally produce feature importance scores as part of how they work. Both build decision trees that track how much each feature helps reduce prediction error [24, 25]. Linear Regression produces coefficients rather than importance scores. The panel regression models also produce coefficients. LightGBM also produces tree-based importance scores, but its rankings were not included, since Random Forest and XGBoost already provided two independent tree-based measures and a third would have weighted this family too heavily in the combined score. Using two different tree-based models rather than just one provided a useful cross-check, if both ranked the same features highly, that would have given more confidence that those features were genuinely important.

The five feature selection methods used were:

1. Pearson and Spearman correlation analysis

Pearson correlation was used to examine linear relationships between each feature and log-transformed electricity demand, while Spearman rank correlation was used to identify monotonic but not necessarily linear relationships [16]. Both measures were computed and used as separate scores in the consolidated ranking.

2. Mutual information

Mutual information measures the statistical dependency between two variables and can capture both linear and nonlinear relationships [12]. It was included as a complementary method to Pearson and Spearman correlation because it does not assume any particular form of the relationship between variables. Mutual information scores were computed using the `mutual_info_regression` function from scikit-learn [12].

3. Recursive Feature Elimination with Cross-Validation (RFE CV)

RFECV was used in place of standard RFE because it automatically determines the optimal number of features to retain, rather than requiring the researcher to set a threshold manually [19]. RFECV works by repeatedly removing the least important feature and using cross-validation to measure the impact on model performance. TimeSeriesSplit cross-validation was used inside RFECV to preserve the chronological order of the data and prevent future data from leaking into the training process. A Linear Regression estimator was used as the base model for RFECV.

4. Random Forest feature importance

A Random Forest model was trained on the exogenous features and the importance of each feature was measured by how much it reduced prediction error across all decision trees in the ensemble [24]. Features that consistently appeared near the top of the trees were assigned higher importance scores.

5. XGBoost feature importance

An XGBoost model was trained on the same exogenous feature set, and feature importance scores based on each feature’s contribution to model performance were extracted [25]. XGBoost importance was included as a second tree-based importance measure to check whether the Random Forest and XGBoost methods agreed on which features were most useful.

After all five methods were applied, the scores from each method were normalised to a 0 to 1 scale and averaged to produce a single consolidated ranking score for each feature. The top five exogenous features by consolidated score were selected: `population_rolling_mean_3` (score = 0.774), `scaled_log_population` (score = 0.665), `temp_rolling_mean_3` (score = 0.512), `pop_x_electricity_access` (score = 0.510), and `temp_x_urbanisation` (score = 0.444).

SHAP (SHapley Additive Explanations) analysis was also carried out using the trained Random Forest model to explain how each individual feature contributed to specific predictions [20]. SHAP provides a more detailed picture than feature importance scores alone because it shows both the direction and the magnitude of each feature’s effect on each prediction, rather than just an overall average importance.

3.7.3 Three Feature Sets

Based on the two-stage feature selection process, three structured feature sets were defined for use in model training and comparison:

- **the Lag-only Model (Model A):** Contains only the two lag features (`lag_1_log_electricity_demand`, `lag_2_log_electricity_demand`). This set served as an autoregressive baseline, representing the simplest possible forecasting model that uses only past demand values.
- **the Exogenous-only Model (Model B):** Contains only the five selected climate and socioeconomic features: `population_rolling_mean_3`, `scaled_log_population`, `temp_rolling_mean_3`, `pop_x_electricity_access`, and `temp_x_urbanisation`. This set was used to evaluate how well electricity demand could be forecast using only structural variables, without any information about past demand levels.
- **the Multivariate Model (Model C):** Contains the two lag features plus the five exogenous features (seven features in total). This set represented the full multivariate approach and was used as the main modelling framework.

Comparing the performance of models across these three feature sets allowed the study to

directly measure the forecasting value added by including climate and socioeconomic variables on top of historical demand information.

3.8 Model Training

Six modelling approaches were trained and compared. They were chosen as four model families rather than as six separate models, so that the comparison would show whether this forecasting problem rewards flexibility or simplicity. The families were a linear baseline (Linear Regression), panel econometrics (Fixed Effects and Dynamic Fixed Effects Panel Regression), bagging (Random Forest) and boosting (XGBoost and LightGBM). Each model is described in the subsections below, together with the reason it was included.

Several other well-known forecasting methods were not included.

ARIMA and SARIMA were not used because they are mainly univariate and single-series methods [4]. Using them would have required a separate model for each of the 47 countries, each fitted on about 20 annual observations. They would also not allow the study to include climate and socioeconomic predictors in the same way. This would reduce the ability to assess what these predictors add to electricity demand forecasting. Seasonal terms are also not useful for annual data.

LSTM and other deep learning models were not used because the available sample is too small to support them. The dataset contains 935 country-year observations, with about 20 years per country. These models usually benefit from longer and higher-frequency time series. With this sample size, they could overfit and would also be more difficult to explain to an energy planning audience. Since the study aims to support energy planning decisions, interpretability was important in the final model selection. Deep learning is therefore suggested as future work for studies using monthly or daily data.

Prophet was not used because it is mainly designed for a single time series using trend, seasonality and holiday effects. The study uses annual panel data, where seasonality and holiday effects are not relevant. Prophet also does not provide a direct way to pool information across countries.

Support Vector Regression (SVR) was considered but not included. It provides limited additional benefit compared with Random Forest and the boosting models used in the study. It is also more difficult to interpret and is more sensitive to the scaling of input features.

3.8.1 Linear Regression

Linear Regression was used as the baseline model because it is simple, transparent, and widely used in predictive modelling [16]. It assumes a linear relationship between the input features and the target variable, making it highly interpretable. For Linear Regression, the model coefficients directly show the direction and magnitude of each feature’s effect on electricity demand. This interpretability was an important consideration for a study aimed at supporting energy planning and decision-making by policymakers. Linear Regression was applied using all three feature sets.

3.8.2 Fixed Effects Panel Regression

Fixed Effects Panel Regression was included because the dataset contained repeated observations for 47 countries over 22 years, which is a classic panel data structure. The Fixed Effects

model accounts for country-specific characteristics that do not change much over time, such as geographic location, historical energy infrastructure, and institutional factors, by estimating a separate intercept for each country [18]. This allows the model to isolate the within-country effects of the predictor variables without those estimates being distorted by permanent differences between countries. The model was implemented using the `PanelOLS` function from the `linearmodels` Python library, with `entity_effects=True` to include country fixed effects.

3.8.3 Dynamic Fixed Effects Panel Regression

Dynamic Fixed Effects Panel Regression extends the standard Fixed Effects model by including lagged electricity demand as an explanatory variable. This allowed the model to capture the autoregressive nature of electricity demand, where current demand is partially determined by past demand levels [18]. This model was applied only to feature sets that included lag features (Model A and Model C), since including a lag of the dependent variable in a model without temporal features would not add any information beyond what the static fixed effects model already captured.

3.8.4 Random Forest

Random Forest was included because it is one of the most widely used and consistently strong-performing machine learning models for tabular regression tasks [24]. It builds a large number of decision trees on random subsets of the training data and features, and combines their predictions through averaging. This makes it much less sensitive to noise and less prone to overfitting than a single decision tree. For this study, a Random Forest with 300 trees was used as the default configuration, and hyperparameter tuning was applied as described in Section 3.9.

3.8.5 XGBoost

XGBoost (Extreme Gradient Boosting) was included because it is a highly efficient and flexible gradient boosting algorithm that has shown strong performance across a wide range of forecasting and regression tasks in the literature [25, 26]. Unlike Random Forest, which builds all trees in parallel, XGBoost builds trees sequentially, with each new tree focused on correcting the errors of the previous ones. It also includes built-in regularisation parameters that help to prevent overfitting. Hyperparameter tuning was applied to the XGBoost model, as described in Section 3.9.

3.8.6 LightGBM

LightGBM [15] is a gradient boosting framework developed by Microsoft that uses a leaf-wise tree growing strategy, as opposed to the depth-wise strategy used by XGBoost. This approach can produce more accurate models in fewer iterations, particularly on larger datasets, although it can also be more prone to overfitting if not properly tuned. LightGBM was included to explore whether this newer approach offered any additional benefits over XGBoost in the specific context of this study. Hyperparameter tuning was also applied to LightGBM.

3.9 Hyperparameter Tuning

Hyperparameter tuning was performed for Random Forest, XGBoost, and LightGBM to find the best model settings for each combination of model and feature set. Grid search with TimeSeriesSplit cross-validation was used for all three models. TimeSeriesSplit was chosen over standard k-fold cross-validation because it preserves the chronological order of the data, preventing future information from being used during cross-validation [13].

Tuning was performed once per model-feature set combination using only the training and validation period data (2000 to 2017). The 2018 to 2021 test set was never used at any stage of the tuning process. The hyperparameter grids used were as follows:

For Random Forest: number of trees (100, 200, 300), maximum tree depth (5, 10, 15, or no limit), and minimum samples required to split a node (2, 5, 10).

For XGBoost: number of trees (100, 200, 300), maximum tree depth (3, 6, 9), learning rate (0.01, 0.05, 0.1), L1 regularisation (0, 0.1, 0.5), and L2 regularisation (1.0, 1.5, 2.0).

For LightGBM: number of trees (100, 200, 300), maximum tree depth (3, 6, 9), learning rate (0.01, 0.05, 0.1), and number of leaves (15, 31, 63).

The best hyperparameters identified through grid search were then used for all walk-forward validation iterations for that model-feature set combination.

3.10 Model Evaluation

The forecasting models were evaluated using four standard metrics that are widely used in the electricity demand and energy forecasting literature [16, 37].

Mean Absolute Error (MAE) measures the average absolute difference between predicted and actual values:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

where y_i represents the actual electricity demand values and $\hat{y}_i$ represents the predicted values. MAE is easy to interpret because it is expressed in the same units as the target variable [37].

Root Mean Squared Error (RMSE) measures the square root of the average squared difference between predicted and actual values:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

RMSE gives more weight to large errors than MAE does, which is important in electricity demand forecasting because large forecasting errors can have serious consequences for infrastructure planning [12].

Mean Absolute Percentage Error (MAPE) expresses the forecasting error as a percentage of the actual values:

$$MAPE = \frac{100}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

MAPE is useful for comparing performance across different models and feature sets because it scales the error relative to the size of the actual demand, making results interpretable regardless of the scale of the target variable [13].

Coefficient of Determination (R^2) measures how well the model explains the variation in the target variable:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

where $\bar{y}$ is the mean of the observed electricity demand values. An R^2 value of 1.0 means the model explains all of the variation in the data, while a value of 0 means the model does no better than simply predicting the mean every time [38].

In addition to forecasting accuracy, model interpretability was also considered during model selection. Linear Regression was considered more interpretable than the ensemble and boosting methods because its coefficients can be read directly as estimates of each feature’s effect on demand.

3.11 Training, Validation, and Data Integrity

3.11.1 Train-Test Split

A chronological train-test split was applied to prevent any future information from influencing the model training process. All observations from 2002 to 2017, 747 rows, were used for model training and walk-forward validation. Observations from 2018 to 2021 were held out as an independent test set and were not used at any stage of model training, feature selection, or hyperparameter tuning. This four-year test set provided a meaningful evaluation of how well the models could generalise to completely unseen data, reflecting a realistic forecasting scenario where past data is used to predict future demand [13, 31].

Table 3.4: Data Splits Used for Model Development and Evaluation

Split	Years	Rows	Countries	Purpose
Training and validation	2002-2017	747	47	Walk-forward validation, hyperparameter tuning, feature selection, and scaler fitting.
Test	2018-2021	188	47	Single final evaluation; used once.
Total modelling frame	2002-2021	935	47	-

The modelling frame begins in 2002 rather than 2000 because the two-year lag and the three-year rolling averages have no prior years to draw on for 2000 and 2001. Those rows were dropped rather than imputed, since filling them would have meant inventing demand history.

One further point should be noted for transparency. After the final model was selected and evaluated on the held-out test set, it was refitted on the complete 2002 to 2021 period before being saved for the dashboard. This refit is standard practice for deployment, since a tool intended for use should draw on all available data, but it means the dashboard model is not the same fitted object that produced the reported test metrics. All accuracy figures in Chapter 4 come from the model fitted on training data only.

3.11.2 Walk-Forward Validation

Within the 2002 to 2017 training and validation period, walk-forward validation with an expanding window was applied. The initial training window covered 2002 to 2010, 418 rows, and models were used to predict electricity demand for 2011. The training window was then expanded by one year, models were retrained on data from 2002 to 2011, and predictions were made for 2012. This process continued until predictions were made for 2017, giving seven validation iterations in total.

Walk-forward validation was chosen over standard k-fold cross-validation because it is specifically designed for time series data [13, 32]. Standard k-fold cross-validation randomly assigns observations to folds, which means that future data can end up in the training set and past data can end up in the validation set. This kind of data leakage makes cross-validation results appear better than they actually are, because the model effectively gets to see the future. Walk-forward validation avoids this problem by always training on the past and predicting the future.

An expanding window was preferred over a rolling window because electricity demand forecasting benefits from using as much historical data as possible. As more years of data became available, the models could learn from longer trends in electricity demand, population growth, GDP changes, and climate patterns [31].

One important practical detail concerns what value of the lag feature was supplied to the model when making each prediction. At each walk-forward iteration, the model was given the actual observed electricity demand from the previous year, not a value that the model itself had predicted earlier. For example, when predicting electricity demand for 2013, the actual recorded demand for 2012 was used as the lag feature input. The same approach was used on the final test set: when predicting demand for 2018, 2019, 2020, and 2021, the actual observed demand from the preceding year was used as the lag input in each case. This choice is appropriate for evaluating model quality in an annual panel setting, where the previous year’s demand is known before the forecast for the current year is needed. It does however mean that results would differ in a true multi-step-ahead forecasting scenario where past demand values are also unknown.

Table 3.5: Walk-Forward Validation Windows

Iteration	Training window	Predicts	Training rows
1	2002–2010	2011	418
2	2002–2011	2012	465
3	2002–2012	2013	512
4	2002–2013	2014	559
5	2002–2014	2015	606
6	2002–2015	2016	653
7	2002–2016	2017	700

Seven iterations were run for each of the 17 valid model and feature set combinations, giving 119 validation results in total.

3.11.3 Data Leakage Prevention

Data leakage happens when information from the test period somehow finds its way into the training process, even if that was not intended. When this happens, the model’s performance looks better than it really is, because it has effectively already “seen” part of the test data

during training. Preventing leakage is therefore critical for producing honest and reliable results [13].

Several steps were taken throughout this study to prevent data leakage at every stage of the pipeline:

- **Scaling:** The Z-score standardisation scaler was fitted only on training-period rows (2000 to 2017, since scaling was applied before the lag-related rows for 2000 and 2001 were dropped). The mean and standard deviation used for scaling came entirely from training data. These same parameters were then applied to scale the test-period rows, rather than computing separate parameters from the test data [36].
- **Climate anomaly features:** The baseline used to calculate temperature and precipitation anomalies (the country-level long-run average) was computed using only training-period observations (2000 to 2017). Test-period temperatures were not included in the baseline calculation. This prevented any test-period climate information from leaking into training-period anomaly values.
- **Feature selection:** Both stages of feature selection (the full-feature round and the exogenous-only round) were run using only training-period data. Pearson and Spearman correlations, mutual information scores, RFECV rankings, Random Forest importance, and XGBoost importance were all calculated from training rows only. This ensured that the features chosen for the final model were selected without any influence from the test period.
- **Hyperparameter tuning:** Grid search for Random Forest, XGBoost, and LightGBM was performed using TimeSeriesSplit cross-validation on training data only. The 2018 to 2021 test set was never used during tuning.
- **Walk-forward validation:** At each iteration, models were trained on data up to a cutoff year and then evaluated on the following year only. No future data was ever visible during training.
- **Permutation importance:** A separate copy of the final model was fitted on training data only (2002 to 2017). The permutation test was then run on the genuinely unseen test period (2018 to 2021). This ensured that the importance scores reflected the model's behaviour on data it had truly never seen.
- **Rolling average and variability features:** These used a strictly backward-looking trailing window (the current year and the two years before it). No future values were included in any rolling calculation. For training-period rows, the window contained only training-period data.
- **Missing value interpolation:** Forward-only interpolation and forward fill were used to handle missing values, meaning that gaps were filled using past values only. No future values were pulled backward to fill earlier gaps.

3.12 Visualization and Interpretation

An interactive decision-support dashboard was developed using Streamlit [34]. The dashboard was designed for one primary user, with two secondary ones. The primary user is an energy planning officer working in a national ministry of energy, a national utility or a regulatory authority in a Sub-Saharan African country. This user has a policy or engineering background rather than a data science one, works with demand figures regularly, and needs to see how

their own country’s demand has moved, what has moved with it, and how a forecast performed against what actually happened. They will not train a model and do not want to. The Country Explorer and Forecasting pages were built for this user, which is why every figure on those pages is shown in terawatt-hours rather than log units, and why the model is loaded pre-trained so that nothing has to be run.

The first secondary user is a policy analyst in a development organisation or a regional body working on cross-country comparison. This user is served by the Overview page, which places all 47 countries on the same axes.

The second secondary user is a researcher who wants to inspect the model itself. The Model Insights page serves this user, showing the final coefficients, the three-way feature set comparison and the full walk-forward table for all 17 combinations. This page is deliberately the last of the four, because it is the only one that assumes familiarity with the modelling.

Two limits on use follow from this. The dashboard does not forecast beyond 2021, because doing so would need projected population and temperature values that this study did not produce. And it is a national and regional tool, not an operational one: annual country totals cannot support decisions about dispatch, network operation or any question below the national and annual level.

The dashboard had four main sections. The Overview page showed total electricity demand trends across all 47 countries, the top countries by demand, and pooled trends in key indicators. The Country Explorer page allowed users to select individual countries or groups of countries and explore their electricity demand trends alongside climate and socioeconomic indicators. The Forecasting page showed actual versus predicted electricity demand for any selected country, with accuracy metrics displayed for the test period (2018 to 2021). The Model Insights page showed the final model’s feature coefficients, the three-way feature set comparison, and the full walk-forward validation summary table for all models.

The dashboard loaded the pre-trained final model, the saved feature list, and the engineered dataset automatically from saved files, meaning that no model training was required at runtime. All predictions shown in the dashboard were back-transformed from log scale to TWh using the `expm1` function so that the results were expressed in meaningful, interpretable units.

3.13 Methodological Framework

Table 3.6 summarises the full methodological pipeline of the study, from data collection through to dashboard visualisation.

Table 3.6: Summary of the Methodological Framework

Phase	Methodological Activity	Output
1	Collect electricity demand, climate, and socioeconomic datasets from multiple sources.	Raw datasets
2	Integrate datasets, clean the data, treat missing values, detect outliers, assess normality, and standardise predictor variables.	Preprocessed panel dataset
3	Construct engineered features, including logarithmic, lag, growth, interaction, trend, anomaly, rolling, and variability features.	Engineered feature dataset
4	Perform two-stage feature selection using statistical and machine learning techniques.	Selected feature sets
5	Train Linear Regression, Fixed Effects, Dynamic Fixed Effects, Random Forest, XGBoost, and LightGBM forecasting models.	Trained forecasting models
6	Optimise machine learning models using grid search and time-series cross-validation.	Optimised model parameters
7	Evaluate models using walk-forward validation and forecasting performance metrics.	Validation results
8	Assess final model performance using an independent test dataset and compare feature sets.	Best-performing model
9	Assess predictor importance using permutation importance.	Feature importance rankings
10	Develop an interactive Streamlit dashboard for visualising forecasts and model outputs.	Decision-support dashboard

3.14 Ethical Considerations

This study used publicly available country-level datasets obtained from trusted international sources. The study did not involve human participants, personal information, or sensitive individual-level records. The risks related to privacy, confidentiality, and informed consent were therefore minimal. All data sources were properly acknowledged and cited, and the study followed accepted academic research practices.

3.14.1 Data Integrity and Responsible Use

All datasets were used only for academic and research purposes. Every step of the data cleaning, integration, and modelling process was documented to maintain consistency and allow for reproducibility. Findings were presented as they emerged from the analysis, without selective reporting or manipulation.

3.14.2 Fair Representation of Countries

The same methodology was applied consistently across all 47 countries to ensure that no country was treated differently in terms of data processing, modelling, or evaluation. Results were interpreted with awareness of the differences in data quality, infrastructure, and economic context that exist across the region.

3.14.3 Transparency and Explainability

All steps in the forecasting pipeline, including the data used, the features selected, the models trained, and the validation procedures applied, were fully documented and explained. Feature importance analysis using Random Forest importance, XGBoost importance, SHAP values, and permutation importance was used to make the models more transparent and their predictions more understandable.

3.14.4 Avoidance of Misuse

The forecasting models and dashboard developed in this study were intended solely to support analysis and decision-making. The limitations of the models and the factors that were not included in the analysis were clearly communicated alongside the results, so that anyone using the findings was aware of what the models could and could not do.

3.14.5 Data Sovereignty and Publication of Results

Although this study used publicly available secondary data that did not involve personal or sensitive records, data sovereignty considerations are relevant when conducting research involving African countries. Data sovereignty refers to the principle that data collected from or about a country should be governed in accordance with that country's laws and interests [39]. A growing number of African countries are introducing data protection frameworks that address how data about their citizens, economies, and environments is collected, stored, shared, and used in research [39].

Even though the datasets used in this study were made available by international organisations for public use, the data from each country was treated with care and respect. Country-level results were presented alongside the known limitations of the data to ensure that findings were not taken out of context or applied in ways that could misrepresent any country's energy situation. The study also made sure that all data sources were clearly stated and that results were shared in a way that is open and accessible, so that researchers and policymakers in Sub-Saharan Africa could also benefit from the work.

3.15 Methodological Constraints

Although the methodology was designed as carefully as possible, several constraints should be acknowledged. The use of annual data was the biggest limitation, as it meant the models could not detect seasonal patterns or short-term fluctuations in electricity demand. Monthly or daily data would have been more informative for this purpose but was not consistently available across all countries [7, 13].

The reliance on secondary datasets from multiple sources also introduced some risk of inconsistency. Differences in how variables were defined and measured across different countries and different source organisations may have introduced noise into the integrated dataset, even after careful preprocessing [17].

The exclusion of variables like electricity prices, energy policies, and consumer behaviour was also a constraint. These factors can influence electricity demand significantly, but the data needed to include them at the country level was not available in a consistent form [10, 11].

Finally, although walk-forward validation was used to reduce forecasting bias, the relatively short historical period available for some countries (particularly those where data was only

available from around 2005 or later) meant that the initial training windows for those countries were smaller than ideal, which may have affected the stability of their model estimates.

Chapter 4

Results

4.1 Introduction

This chapter presents the results of the data analysis, feature engineering, feature selection, model training, and model evaluation carried out in this study. The chapter is organised to follow the methodological pipeline described in Chapter 3. It begins with the dataset integration and descriptive statistics, moves through the exploratory data analysis and preprocessing results, then covers the feature engineering and selection outcomes, and concludes with the model performance results and final model selection.

4.2 Dataset Integration and Descriptive Statistics

Seven datasets from different sources were successfully merged into a single panel dataset using country ISO code and year as joint identifiers. The initial merged dataset contained 1,056 rows representing 48 Sub-Saharan African countries observed annually over the period 2000 to 2021. South Sudan was kept in the dataset at this point so that its data coverage could be examined alongside all other countries. The exploratory analysis confirmed that South Sudan had complete data only from 2011 onward (the year it became independent), with all variables missing for the entire 2000 to 2010 period. Attempting to impute 11 consecutive years of missing data from just ten years of observations would have produced unreliable values, so South Sudan was dropped after this comparison, leaving 1,034 rows across 47 countries for the rest of the analysis.

Table [4.1](#) presents the descriptive statistics for all variables in the merged dataset. These statistics were calculated on the full 1,056-row dataset before South Sudan was taken out, which is why some of the counts in the table are slightly above 1,034. The counts also differ from one variable to another because a few variables still had missing values at this stage.

Table 4.1: Descriptive Statistics of the Integrated Dataset Before Exclusion of South Sudan (48 countries, 2000-2021)

Variable	Count	Mean	Std Dev	Min	Median	Max
Electricity Demand (TWh)	1,044	8.17	32.49	0.02	1.46	243.66
Avg Temperature (°C)	1,056	24.70	3.33	10.64	25.25	29.82
Precipitation (mm)	1,056	1,045.81	657.83	33.51	1,033.79	3,308.22
Population	1,056	19,167,860	29,637,360	144,018	10,738,710	218,529,300
GDP per Capita (USD)	1,032	1,787.03	2,367.95	109.59	872.49	18,210.56
Urbanisation Rate (%)	1,056	39.73	16.80	8.04	38.26	90.79
Electricity Access (%)	1,040	37.84	24.67	0.80	35.15	100.00
Urban Electricity Access (%)	1,049	63.84	23.13	3.50	67.00	100.00

The descriptive statistics revealed a number of important characteristics of the dataset. Electricity demand showed very high right-skewness, with a mean of 8.17 TWh but a median of only 1.46 TWh, confirming that a small number of large-demand countries (primarily South Africa and Nigeria) pulled the average far above the typical country’s consumption level. GDP per capita showed a similar pattern, with a mean of USD 1,787 but a median of only USD 872. Population ranged from just over 144,000 (small island states like Seychelles) to over 218 million (Nigeria). Temperature ranged from 10.64°C to 29.82°C, reflecting the diversity of climates from cooler highland countries like Lesotho to the very hot arid regions of the Sahel.

4.2.1 Demand Concentration

One of the most important structural findings from the descriptive analysis was the extreme concentration of electricity demand among a small number of countries. Five countries (South Africa, Nigeria, Mozambique, Zambia, and Ghana) together accounted for 73.6% of the total average annual electricity demand across all 47 countries in the dataset. South Africa alone dominated, with an average annual demand of approximately 230 TWh, which was around nine times higher than Nigeria’s average of approximately 25 TWh, the next highest in the region.

Figure [4.1](#) shows the top 15 countries by average electricity demand over the study period.

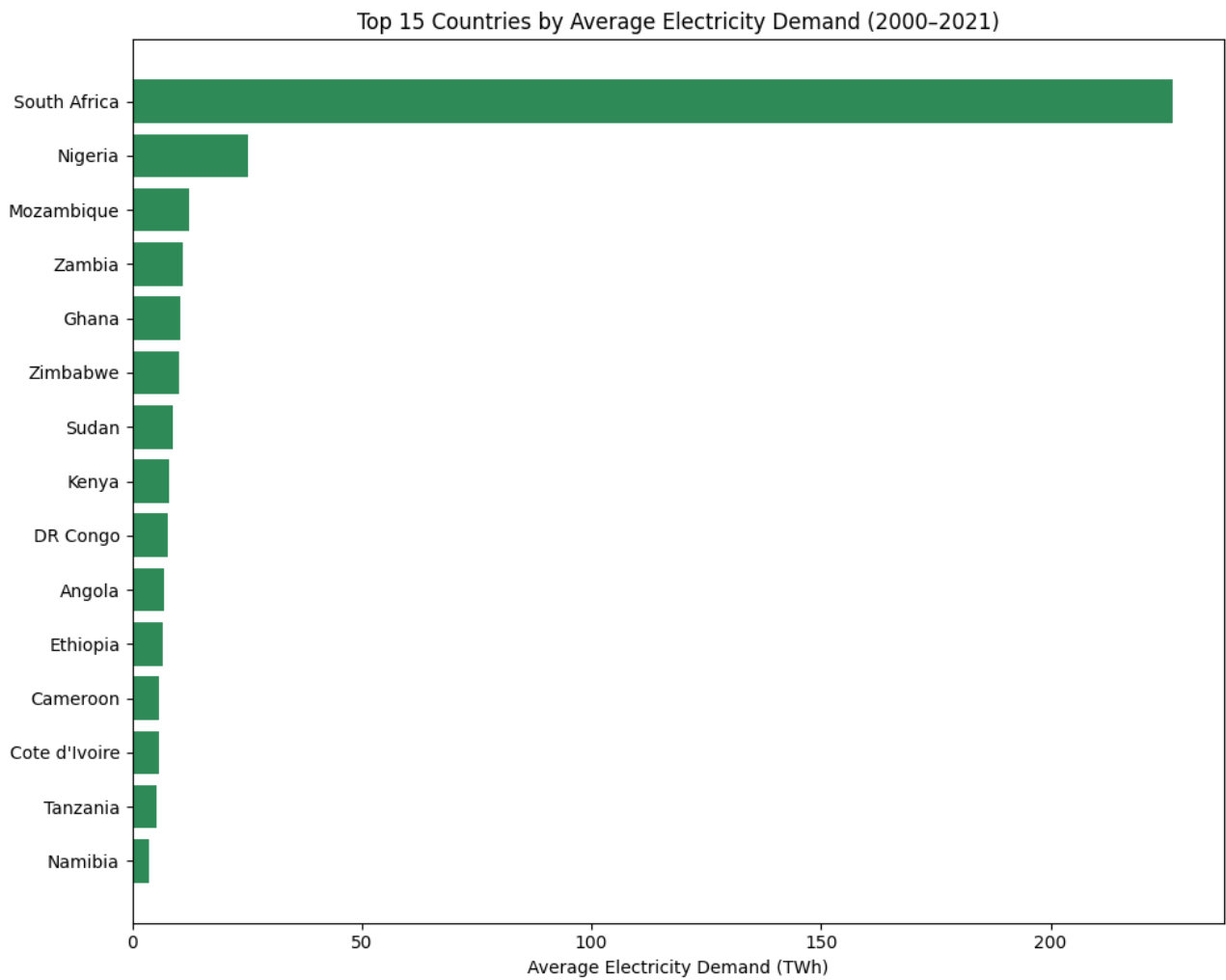


Figure 4.1: Top 15 Sub-Saharan African Countries by Average Annual Electricity Demand (2000-2021)

4.3 Exploratory Data Analysis

4.3.1 Temporal Trends

Figure 4.2 shows the average trend for all 8 variables across the 47 countries over the study period.



Figure 4.2: Average Trends Over Time for All Variables (2000-2021)

The trend plots revealed several important and interesting patterns across the region. Starting with electricity demand, there was a steady and consistent upward trend from around 6 TWh average per country in 2000 to nearly 10 TWh by 2021. This represents roughly a 67% increase over the 22-year period, which is broadly in line with the region's pace of economic and population expansion. The fact that the line is smooth and mostly upward without major dips suggests that while individual countries may have had disruptions, the average across 47 countries was stable enough to absorb those local shocks.

Population, urbanisation rate, electricity access, and urban electricity access all showed the same kind of consistent upward pattern. More people means more households that need power. More people moving into cities means higher per-person electricity use (urban households and businesses use more electricity than rural ones). And rising electricity access means more people are being connected to the grid for the first time, adding directly to total demand. These four trends essentially reinforce each other, which is why they all move in the same direction at roughly the same pace.

GDP per capita tells a slightly more complicated story. It rose strongly from 2000 to around 2014, but then the trend became more volatile and less clear. This is most likely explained by the global commodity price crash of 2014 to 2016, which hit Sub-Saharan Africa particularly hard since many of the region's economies depend heavily on oil, minerals, and agricultural exports. When commodity prices fell sharply, so did GDP in many countries, and the recovery was uneven. This pattern has been documented in the literature on economic cycles in developing

regions [2].

Temperature shows a very different kind of trend. It is mostly flat but with a slight upward drift over the 22 years, and with more visible year-to-year variation than the socioeconomic variables. The slight upward creep in average temperature is consistent with the broader signal of long-term warming seen across the African continent, which has been linked to global climate change [1]. The year-to-year variation reflects the fact that temperatures do not increase steadily but move up and down around a slowly rising baseline.

Precipitation is the most eye catching contrast in the whole set of plots. Unlike every other variable, it shows no clear upward or downward direction at all. It bounces up and down with no visible trend, which tells us something important, rainfall across Sub-Saharan Africa has not been increasing or decreasing in a consistent way over this period. This makes sense given that precipitation in the tropics and subtropics is driven by complex seasonal and interannual climate patterns, which do not produce steady trends in the same way that temperature or population growth does [11]. This is also why the precipitation anomaly feature was created during feature engineering, to capture unusual years rather than any long-term direction.

Figure 4.3 shows the electricity demand trend for six selected countries over the study period.

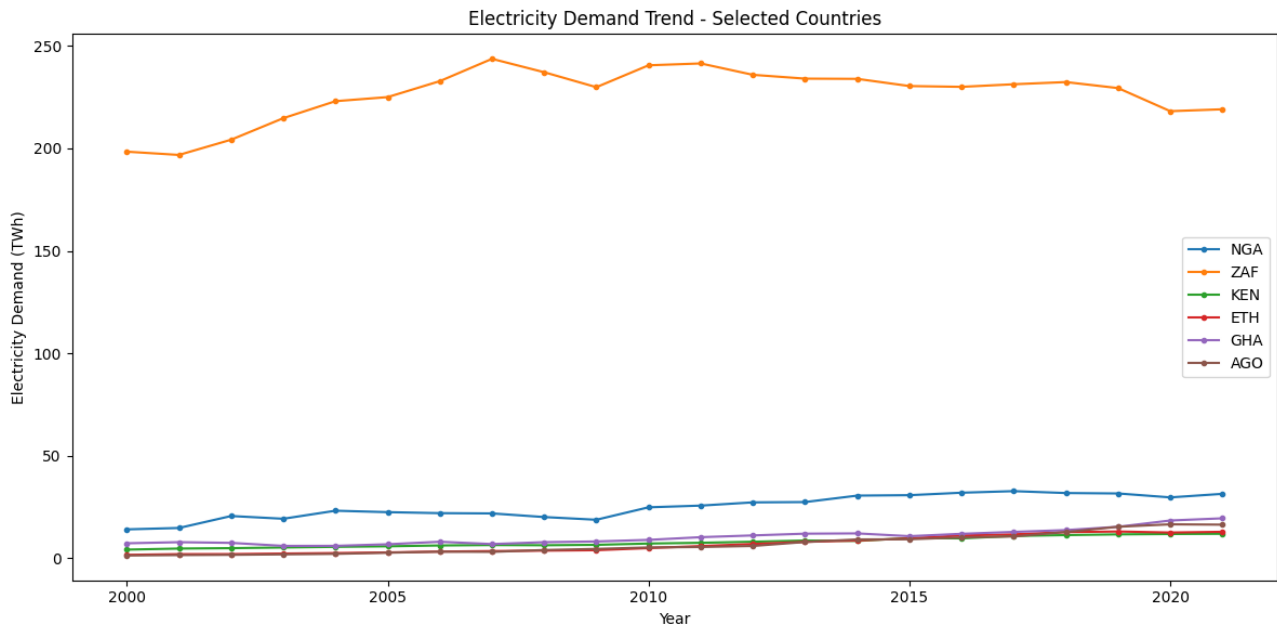


Figure 4.3: Electricity Demand Trends for Selected Sub-Saharan African Countries (2000-2021)

The individual country trends made the huge gap in electricity demand between South Africa and everyone else very clear. South Africa maintained demand levels between 200 and 240 TWh throughout the study period, putting it in a completely different category from all other countries. It is worth pausing to understand why this is the case.

South Africa leads the African continent in electricity generation, and this is directly tied to the structure of its economy. The country has a large, energy-intensive coal mining industry and mostly uses coal to meet its energy needs. The mining sector alone, including mineral processing, consumes around 30% of Eskom’s electricity output [40]. Beyond mining, the industrial sector as a whole accounted for up to 56% of total electricity consumption in South Africa as of 2019. No other country in Sub-Saharan Africa has anything close to this level of industrial electricity demand.

South Africa also has a well-developed electricity network and one of the highest rates of electricity access in sub-Saharan Africa [1, 41]. Eskom, the state-owned electricity utility, supplies about 90% of power generation in South Africa and across the Southern African Power Pool region. This combination of large industrial demand, extensive grid infrastructure, and high access rates is what separates South Africa so dramatically from every other country in this dataset. Its urbanisation rate is also among the highest in the region, and its GDP per capita is the largest in absolute terms among the 47 countries studied, further reinforcing its position as the dominant electricity consumer [41].

Put all of this together and it is not hard to see why South Africa alone accounts for more electricity consumption than the next ten countries combined. This level of concentration is not just a statistical curiosity, it has practical implications for any analysis of Sub-Saharan African electricity demand, because South Africa’s trajectory (including its post-2008 load-shedding crisis) can pull regional averages significantly, even when the underlying trends in other countries are quite different.

The slight decline visible in South Africa’s demand after around 2008 is also worth noting. This was the period when Eskom, South Africa’s national electricity utility, began experiencing severe power shortages due to years of underinvestment in new generation capacity. Rolling blackouts (known locally as “load-shedding”) became increasingly frequent from 2008 onwards, which forced households and businesses to reduce their electricity consumption or switch to alternatives. So the downward movement in South Africa’s demand line was not because the underlying need for electricity fell, but because supply constraints limited how much could actually be consumed.

Nigeria showed a modest upward trend from approximately 14 TWh to 32 TWh, which might seem small given that Nigeria is the most populous country in Africa. The reason is that Nigeria has historically had very poor electricity infrastructure, with a large share of the population depending on private generators rather than the national grid. GDP growth and population growth therefore did not translate into proportional increases in measured grid electricity demand. Kenya, Ethiopia, Ghana, and Angola all showed gradual upward trends from low baselines, with Ghana growing most visibly in the later years of the study period.

4.3.2 Missing Values by Year

Figure 4.4 shows the total number of missing values per year across all variables.

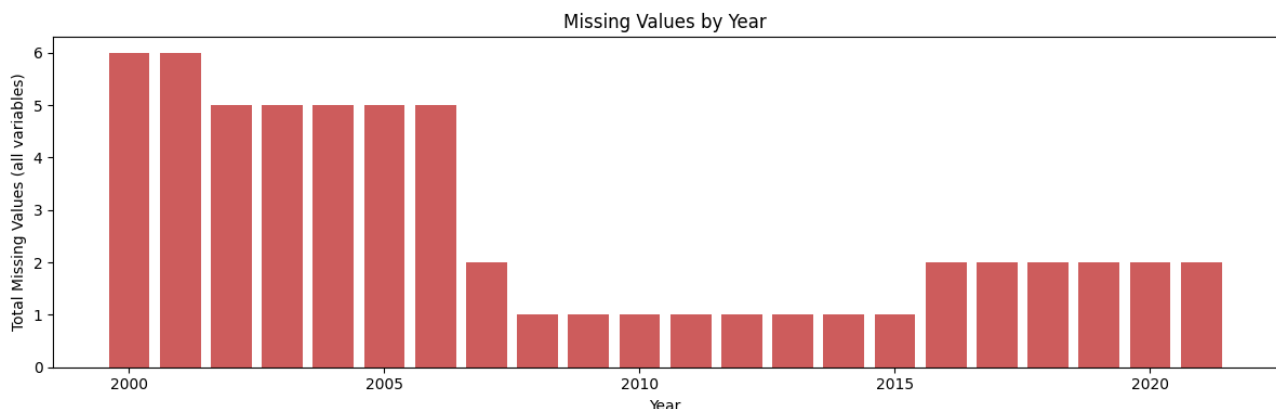


Figure 4.4: Total Missing Values Per Year Across All Variables

Missing values were concentrated in the early years of the study period (2000 to 2006) and

showed a clear declining pattern as reporting improved over time. A small secondary cluster of missing values was visible in the 2016 to 2021 period, mainly due to gaps in the electricity access and urban electricity access data for a small number of countries.

4.4 Preprocessing Results

4.4.1 Normality Test Results

Table 4.2 shows the results of the Shapiro-Wilk normality test and the skewness values for all numeric variables.

Table 4.2: Shapiro-Wilk Normality Test Results

Variable	Shapiro-Wilk Stat	p-value	Normal?	Skewness
Electricity Demand	0.2158	0.000	No	6.44
Average Temperature	0.8979	0.000	No	−1.41
Precipitation	0.9550	0.000	No	0.66
Population	0.5999	0.000	No	3.54
GDP per Capita	0.6321	0.000	No	2.95
Urbanisation Rate	0.9758	0.000	No	0.49
Electricity Access	0.9518	0.000	No	0.54
Urban Electricity Access	0.9623	0.000	No	−0.47

The Shapiro-Wilk test [35] works by comparing the actual distribution of a variable against what a normal distribution would look like. If the two match closely, the test statistic is close to 1 and the p-value is high (above 0.05), meaning there is no strong evidence against normality. If they don't match, the statistic drops below 1 and the p-value falls to near zero. In this study, every variable produced a Shapiro-Wilk statistic well below 1 and a p-value of 0.000, which means all eight variables failed the test at the 5% significance level. The conclusion was clear, none of the variables followed a normal distribution.

The skewness column tells us more specifically what kind of non-normality each variable has. Skewness measures how lopsided the distribution is. A value of zero means the distribution is perfectly symmetric. Positive values mean the right tail is longer (most values are low but a few very high values pull the mean upward), and negative values mean the left tail is longer.

Electricity demand had the highest skewness at 6.44 (strongly positive/right-skewed). This makes a lot of sense when you remember that South Africa's demand of 226 TWh is nine times higher than the next country. Most countries have demand below 10 TWh, but South Africa pulls the upper tail of the distribution out dramatically. Population at 3.54 and GDP per capita at 2.95 showed similar but less extreme right skew, again driven by a small number of outlier countries at the high end.

Average temperature had a skewness of −1.41 (left-skewed), meaning most countries are in the warm tropical range with only a handful of cooler highland countries pulling the lower tail leftward. Precipitation at 0.66 and urbanisation rate at 0.49 were within the acceptable range and required no transformation. Electricity access at 0.54 and urban electricity access at −0.47 were also sufficiently symmetric.

Based on these tests, three variables were selected for log transformation: electricity demand, population, and GDP per capita. The criterion used was an absolute skewness value greater

than 1, which is a widely used rule of thumb in the forecasting literature for identifying variables that benefit from transformation [12].

Figure 4.5 shows the distribution of all variables before transformation.

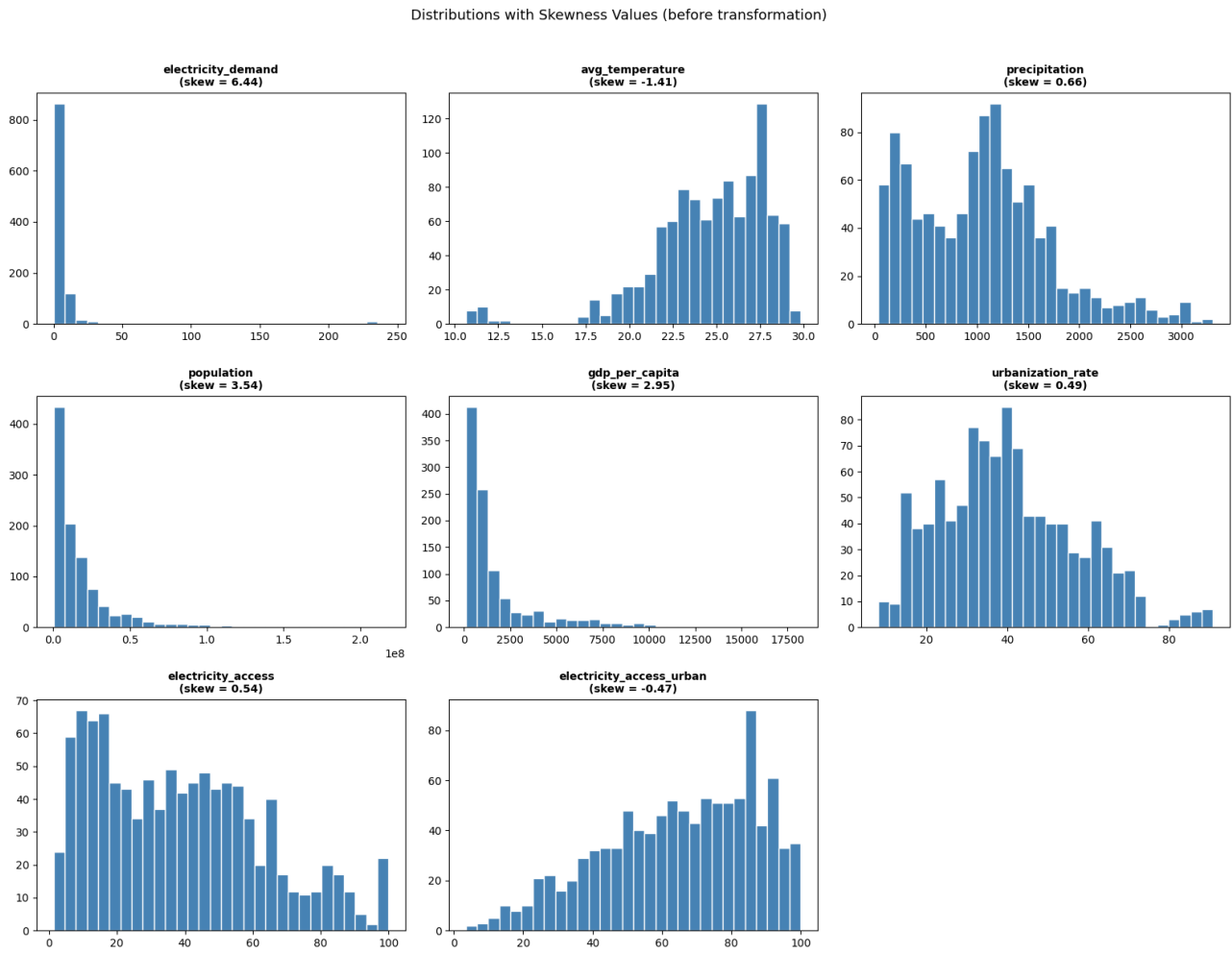


Figure 4.5: Distributions of All Numeric Variables with Skewness Values (Before Transformation)

4.4.2 Outlier Analysis

Figure 4.6 shows the boxplots for all variables before any transformation or treatment.

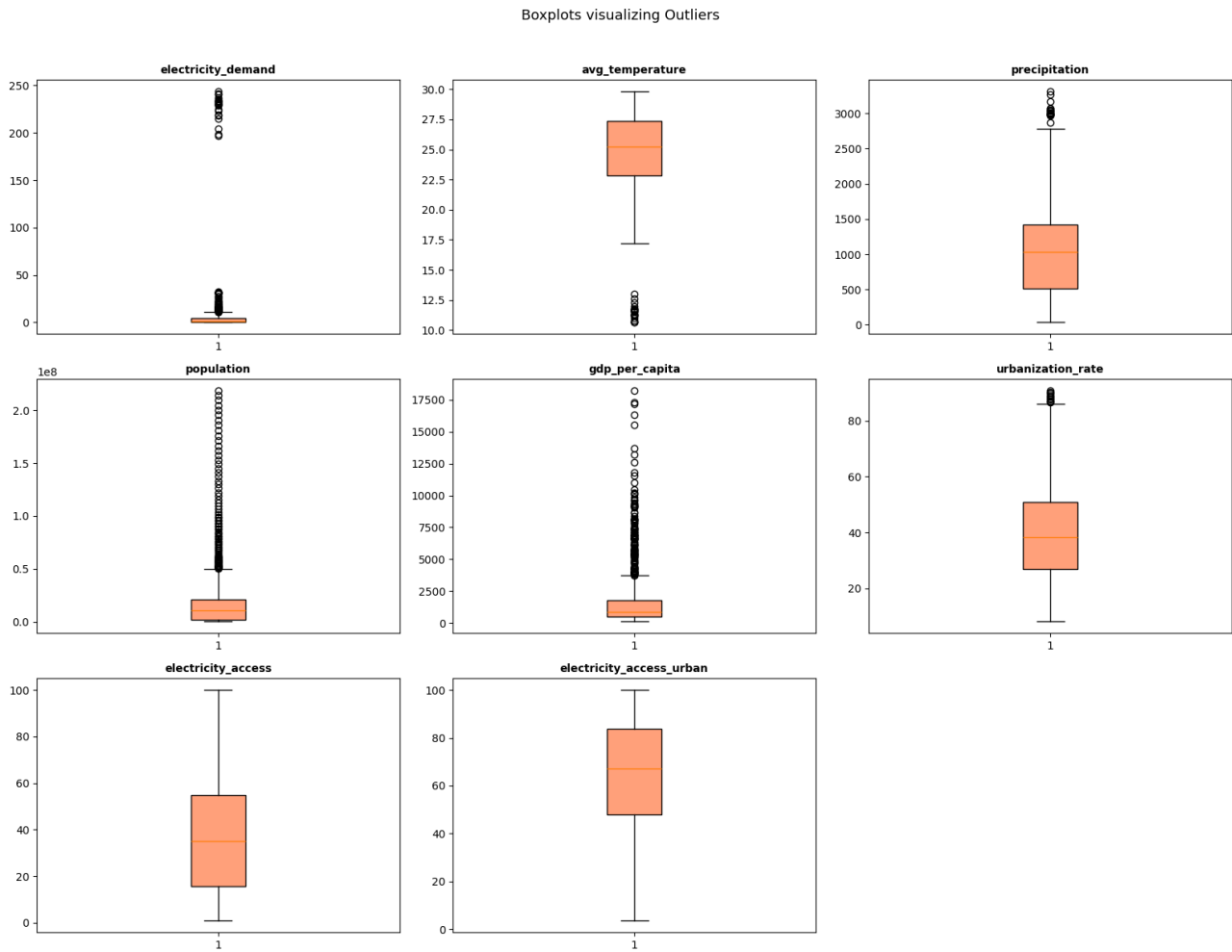


Figure 4.6: Boxplots Showing Distribution and Outliers (genuine extreme values) for All Variables

The boxplots confirmed the findings from the normality tests. Electricity demand, population, and GDP per capita all showed a large number of data points above the upper whisker ($1.5 \times \text{IQR}$), reflecting the presence of high-demand, large-population, or high-income countries like South Africa, Nigeria, and Mauritius. Temperature showed a small number of low-end outliers, representing cooler highland countries like Lesotho and Ethiopia. Electricity access and urban electricity access showed no outliers, confirming that these variables were evenly distributed within their 0-100% range.

As described in Chapter 3, no outlier removal or capping was applied. The values flagged as outliers were genuine reflections of real country-level differences rather than data errors, and removing them would have distorted the dataset and reduced the model's ability to learn from the full diversity of the region.

4.4.3 Missing Value Treatment

After linear interpolation and forward fill were applied, all missing values were successfully resolved with no rows removed from the dataset. To check whether the imputation changed the shape of the data in any meaningful way, the mean, variance, and standard deviation of the affected variables were compared before and after treatment. All three summary statistics changed by less than 1% for every variable. The variance, which measures the overall spread of values around the mean, stayed almost identical, indicating that the filled-in values were

consistent with the surrounding data and did not introduce any unusual spikes or artificial smoothing. The standard deviation (which is just the square root of the variance and is reported in the same units as the original variable) also remained stable. These checks confirmed that linear interpolation preserved the distributional properties of the data reasonably well, and that the imputation did not distort the dataset in a way that would affect the modelling [17].

4.5 Feature Engineering Results

4.5.1 Log Transformation

Figure 4.7 shows the distributions of electricity demand, population, and GDP per capita before and after log transformation.

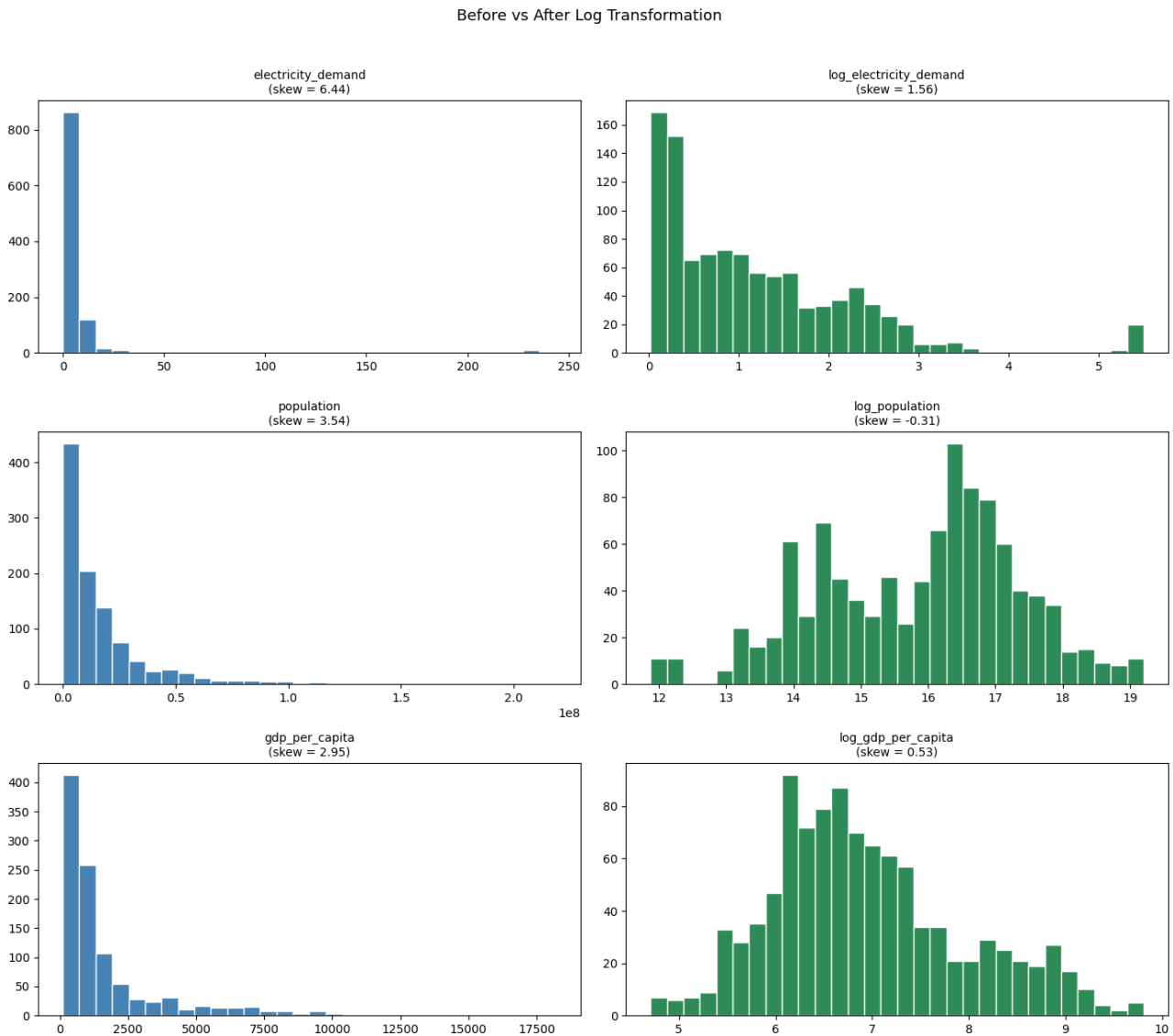


Figure 4.7: Distributions Before and After Log Transformation for Three Skewed Variables

The transformation used was $\log_{1p}(x) = \log(1 + x)$, which is a variant of the natural logarithm (base $e \approx 2.718$). The reason for using $\log_{1p}$ rather than the standard natural log is that $\log(0)$ is undefined, and some countries in the dataset had very small electricity demand values close to zero. Adding 1 before taking the log ensures the calculation is always safe: $\log(1 + 0) = 0$,

so even a zero value maps to a defined output. For larger values the difference between $\log(x)$ and $\log(1 + x)$ is very small and practically irrelevant.

The transformation worked by compressing the scale. A country with demand of 1 TWh gets a log value of $\log(1 + 1) = 0.693$. A country with demand of 10 TWh gets $\log(11) = 2.398$. A country with 230 TWh (South Africa) gets $\log(231) = 5.442$. Without the transformation, South Africa's value is 230 times larger than the 1 TWh country. After transformation, it is only about 8 times larger. The gap is still there, but the scale is compressed enough that the linear model can work with it properly.

The results were clear. Electricity demand's skewness reduced from 6.44 to 1.56, population's from 3.54 to -0.31 , and GDP per capita's from 2.95 to 0.53. All three moved into a much more symmetrical range. Population's skewness actually went slightly negative after transformation (-0.31), meaning it became very slightly left-skewed, but this is a minor issue and well within the acceptable range. The histograms in Figure 4.7 confirm visually that the bell-shaped, more symmetrical distributions emerged after the transformation.

4.5.2 Feature Engineering Summary

In total, 27 features were created across nine engineering categories. These included log-transformed versions of two predictor variables, two lag features, five growth rate features, three interaction features, one trend feature, two climate anomaly features, four rolling average features, one temperature variability feature, and seven scaled predictor variables. Electricity demand was also log-transformed, but as the target variable it is counted separately from the predictors. A full list of all engineered features was provided in Table 3.2 of Chapter 3.

4.6 Feature Selection Results

4.6.1 Stage 1: Full Feature Evaluation and Lag Consistency Check

Figure 4.8 shows a Pearson correlation heatmap covering 17 of the engineered features together with the target, giving 18 variables in total. The heatmap includes the seven scaled predictors, the two lag features, the five growth rate features and the three interaction features. The trend, anomaly, rolling average and variability features are not shown here, but all of them were included in the wider selection pool described in Chapter 3.

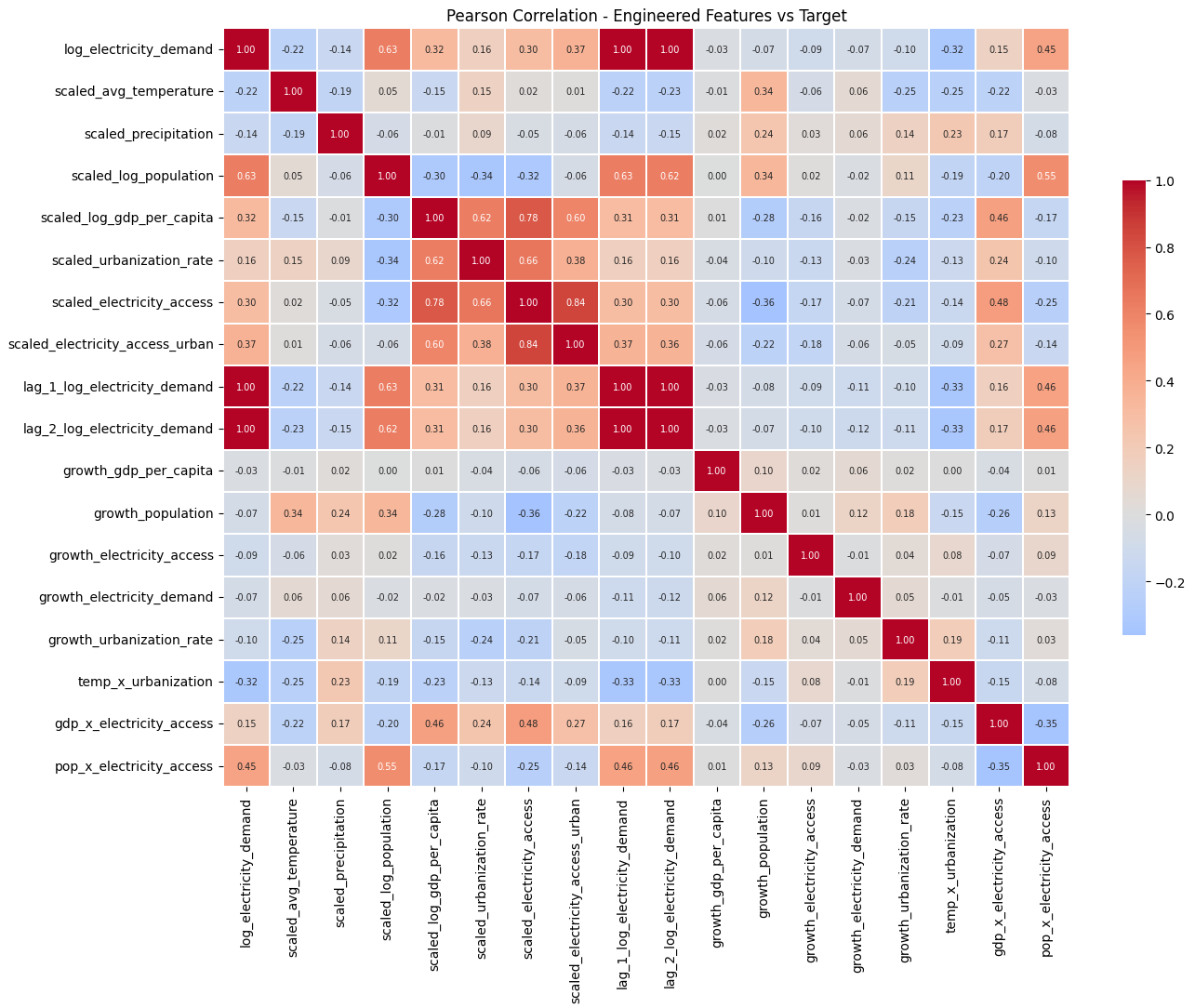


Figure 4.8: Pearson Correlation Heatmap of Scaled, Lag, Growth Rate and Interaction Features with Log-Transformed Electricity Demand

The correlation analysis confirmed that the two lag features dominated the full feature set. The one-year lag (`lag_1_log_electricity_demand`) had a Pearson correlation of +0.999 with the target variable, and the two-year lag had a correlation of +0.997. These were by far the strongest correlations in the dataset, reflecting the highly autoregressive nature of annual electricity demand.

Table 4.3 summarises how the two lag features ranked across the four feature selection methods used in Stage 1.

Table 4.3: Lag Feature Consistency Check Across All Selection Methods

Feature	Pearson	RFECV	RF Importance	XGBoost	Conclusion
lag_1_log_electricity_demand	1st	1st	1st (0.721)	2nd (0.428)	Retained
lag_2_log_electricity_demand	2nd	2nd	2nd (0.170)	1st (0.531)	Retained

Both lag features ranked consistently in the top two positions across all four methods. Lag_1 was ranked first by Pearson correlation, RFECV, and Random Forest importance. Lag_2 was

ranked first only by XGBoost. This consistent pattern across methods confirmed that both lag features were genuinely important and should be retained.

Figure 4.9 shows the RFECV ranking chart for all features.

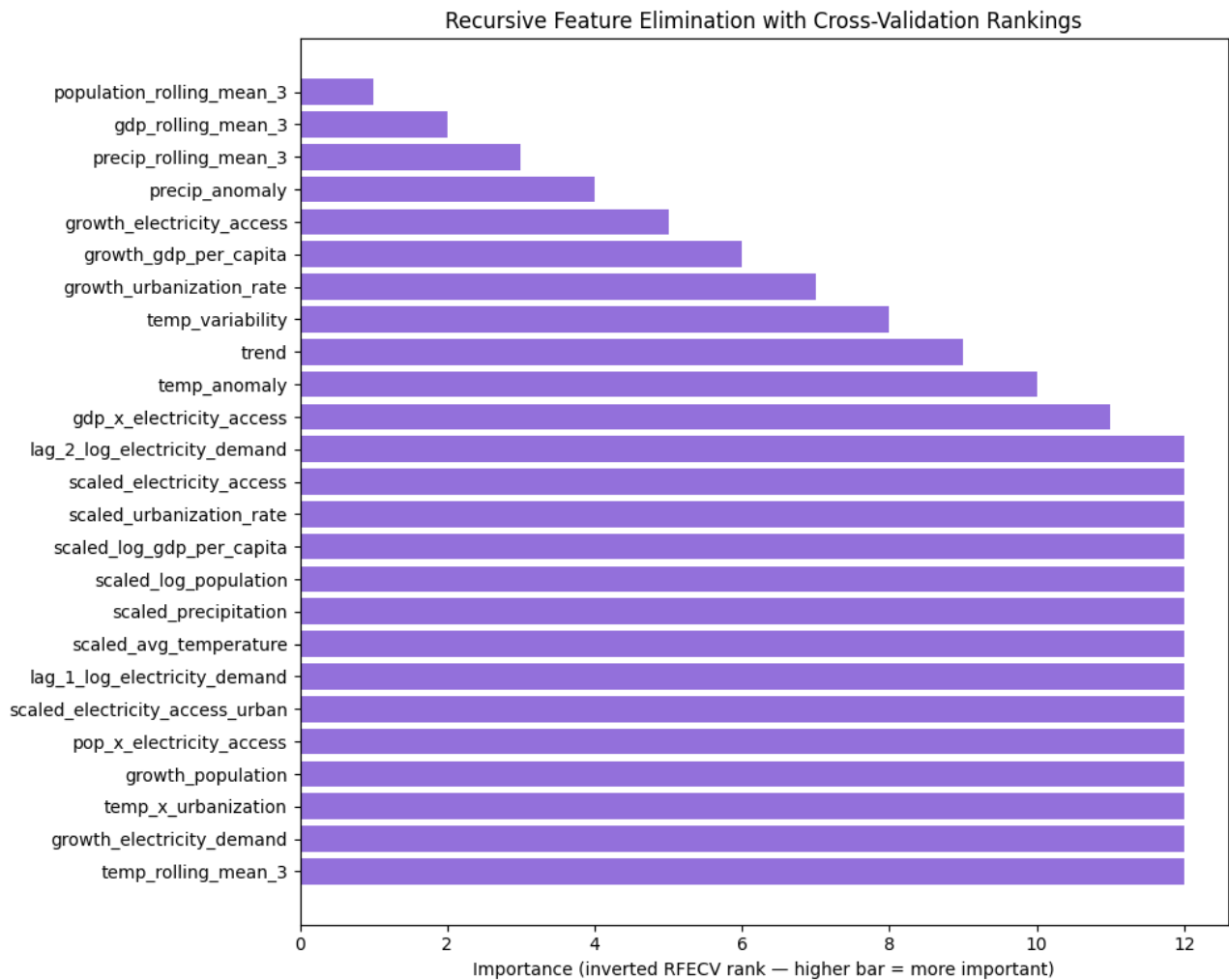


Figure 4.9: RFECV Feature Rankings - All Features (Higher Bar = More Important)

Figure 4.10 shows the importance scores from both models side by side, so it is easy to see where they agree and where they differ.

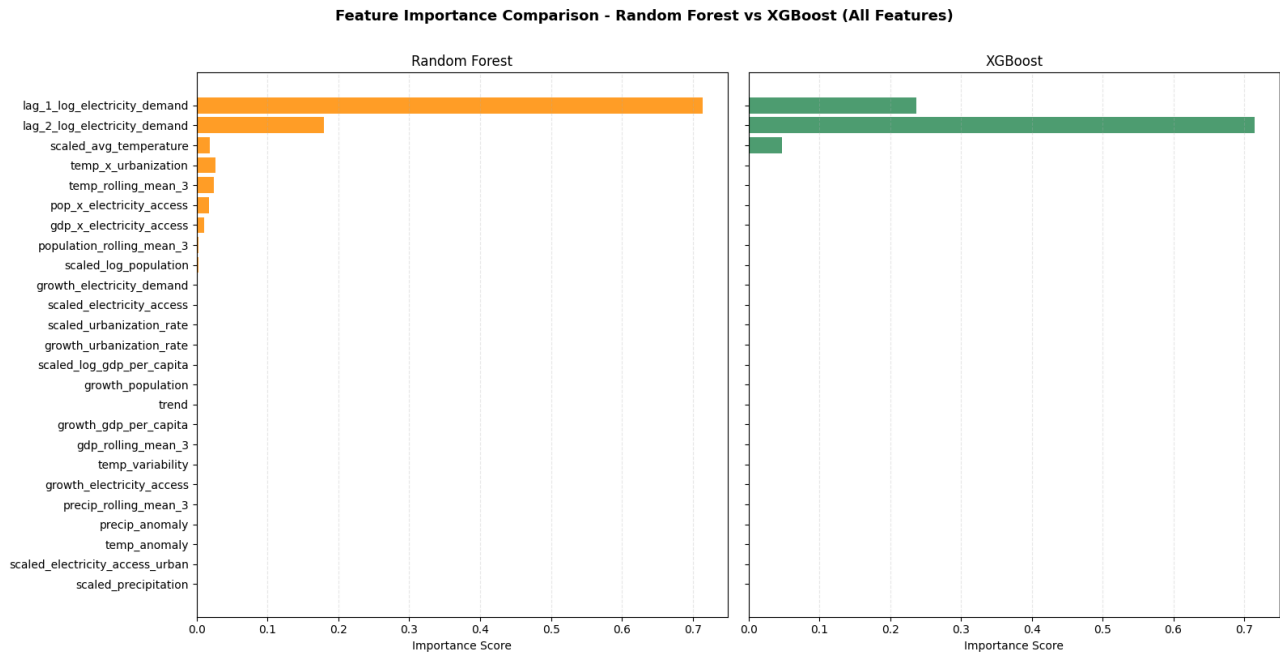


Figure 4.10: Feature Importance Comparison - Random Forest vs XGBoost (All Features)

Both models agree very strongly that the two lag features are by far the most important. Together they account for roughly 90% of all importance in both models. After the lags, average temperature shows up as the next most useful feature in both, though with a much smaller score.

One small difference between the two is that Random Forest ranks lag_1 higher, while XGBoost puts lag_2 slightly ahead.

Figure 4.11 shows the SHAP Feature Attribution plot for the Random Forest model trained on all features.

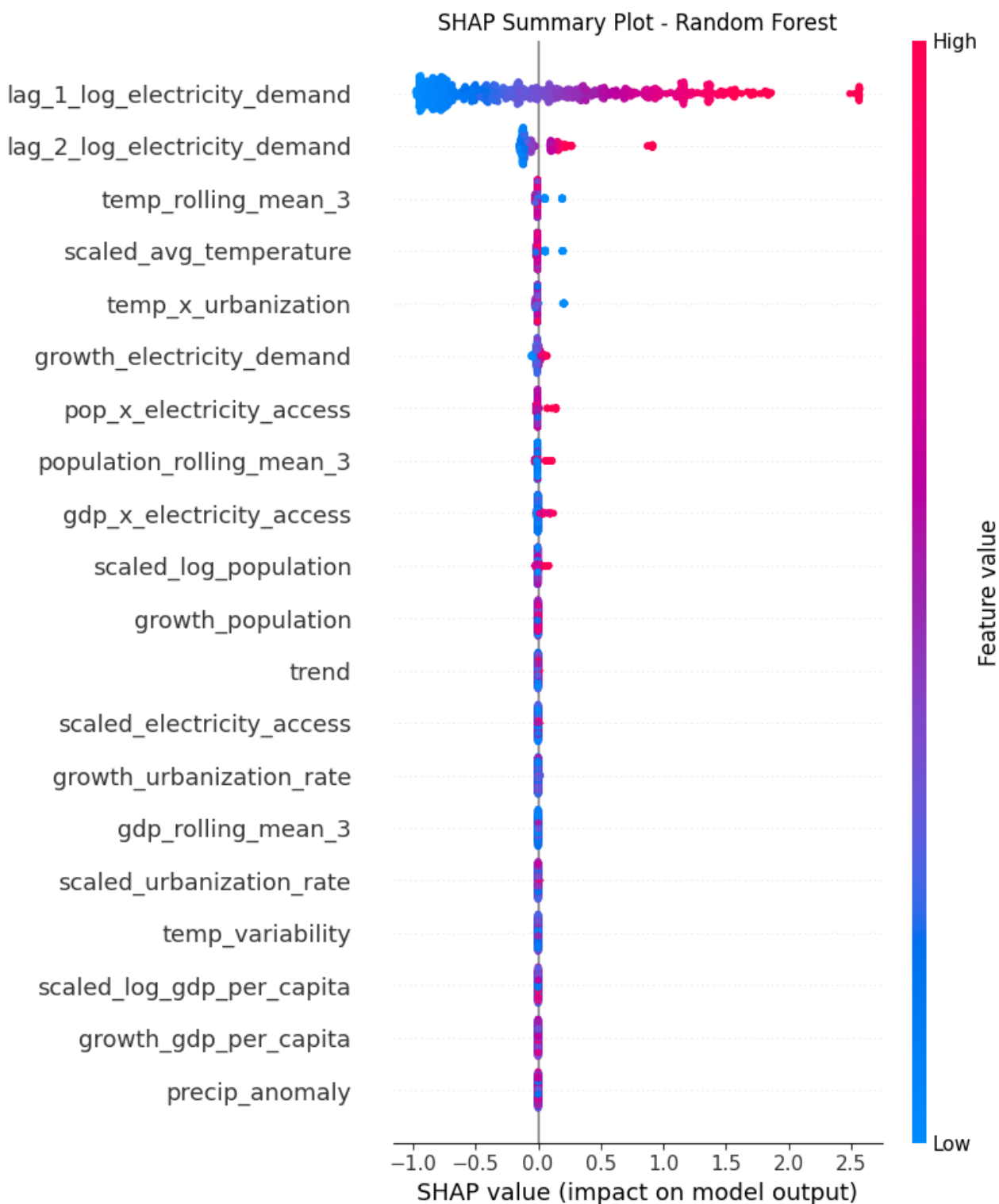


Figure 4.11: SHAP Feature Attribution Plot for the Random Forest Model on all Features

The SHAP analysis confirmed the dominance of lag_1 in the full feature set. In the plot, each row is one feature, each dot is one country-year observation, and the colour shows the value of that feature (red = high, blue = low). The position of each dot on the x-axis shows how much that feature pushed the prediction higher (right) or lower (left) for that observation.

For lag_1, the dots spread widely across the x-axis. Red dots (high lag_1 values, meaning high last-year demand) cluster on the right, pushing predictions up. Blue dots (low lag_1 values) cluster on the left, pushing predictions down. This is exactly what you would expect for a

strong autoregressive signal, countries with high demand last year are predicted to have high demand this year, and countries with low demand last year are predicted to have low demand. The spread is also extremely wide compared to all other features, indicating the sheer scale of lag_1's influence relative to everything else.

All other features had their dots clustered almost entirely at zero on the x-axis. This is a critical finding. It is not just that the other features had small effects, it is that they had virtually no effect at all once the lag feature was present in the model. The model simply did not need them. This is the visual confirmation of what the permutation importance analysis also showed, shuffling the climate and socioeconomic features caused almost no change in model performance, because the model was not really using them.

4.6.2 Stage 2: Exogenous Feature Selection

After excluding the lag features, all five selection methods were applied to the 19 remaining exogenous features (after four features were removed due to multicollinearity). Figure 4.12 shows the correlation of the exogenous features with log-transformed electricity demand.

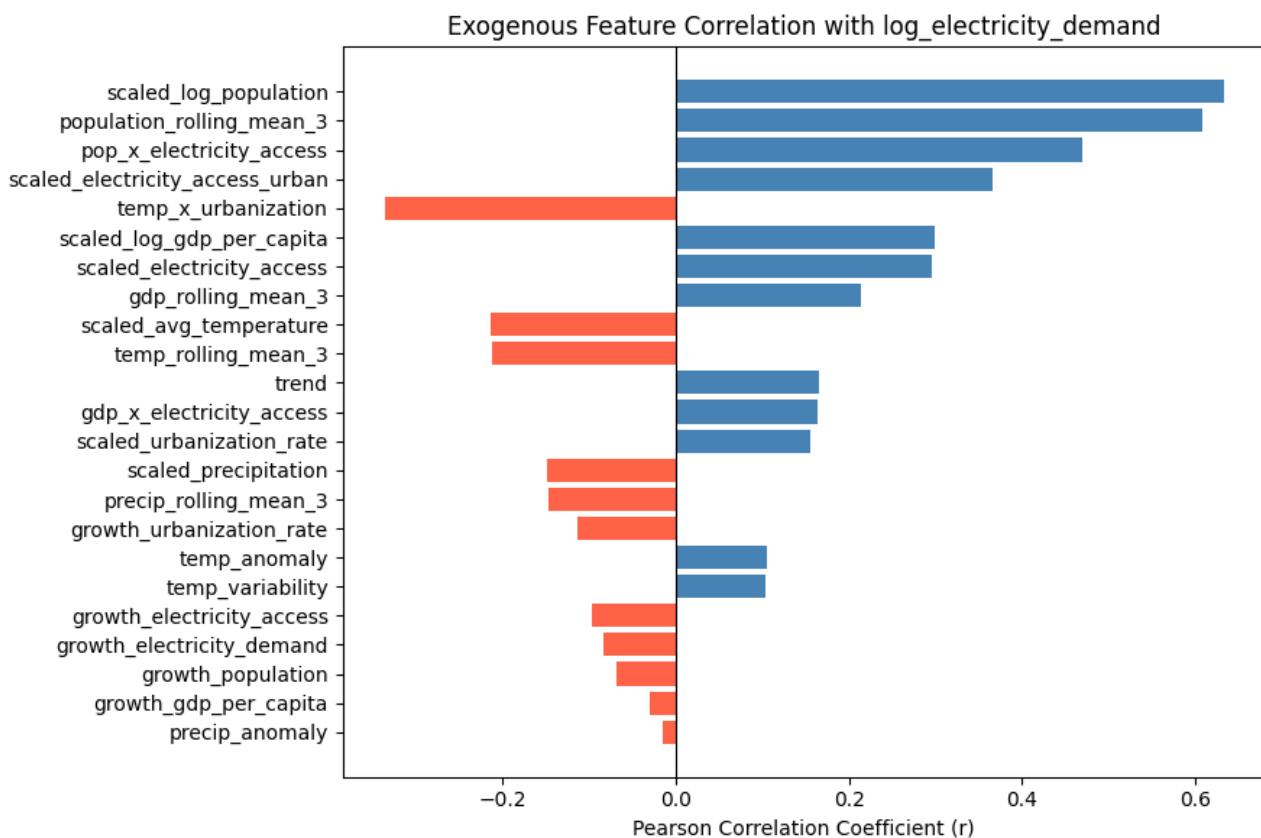


Figure 4.12: Pearson Correlation of Exogenous Features with Log-Transformed Electricity Demand

Without the lag features dominating, the exogenous features showed a much more varied and informative range of correlation values. Scaled log population had the highest positive correlation (0.621) with electricity demand, followed by the population rolling mean (0.606) and the population-electricity access interaction (0.421). Urban electricity access also showed a moderate positive correlation (0.360). The temperature-urbanisation interaction had the strongest negative correlation (-0.334), followed by the temperature rolling mean (-0.229) and average temperature (-0.228). Growth rate features showed weak correlations close to zero.

Figure 4.13 shows the RFECV rankings for the exogenous features.

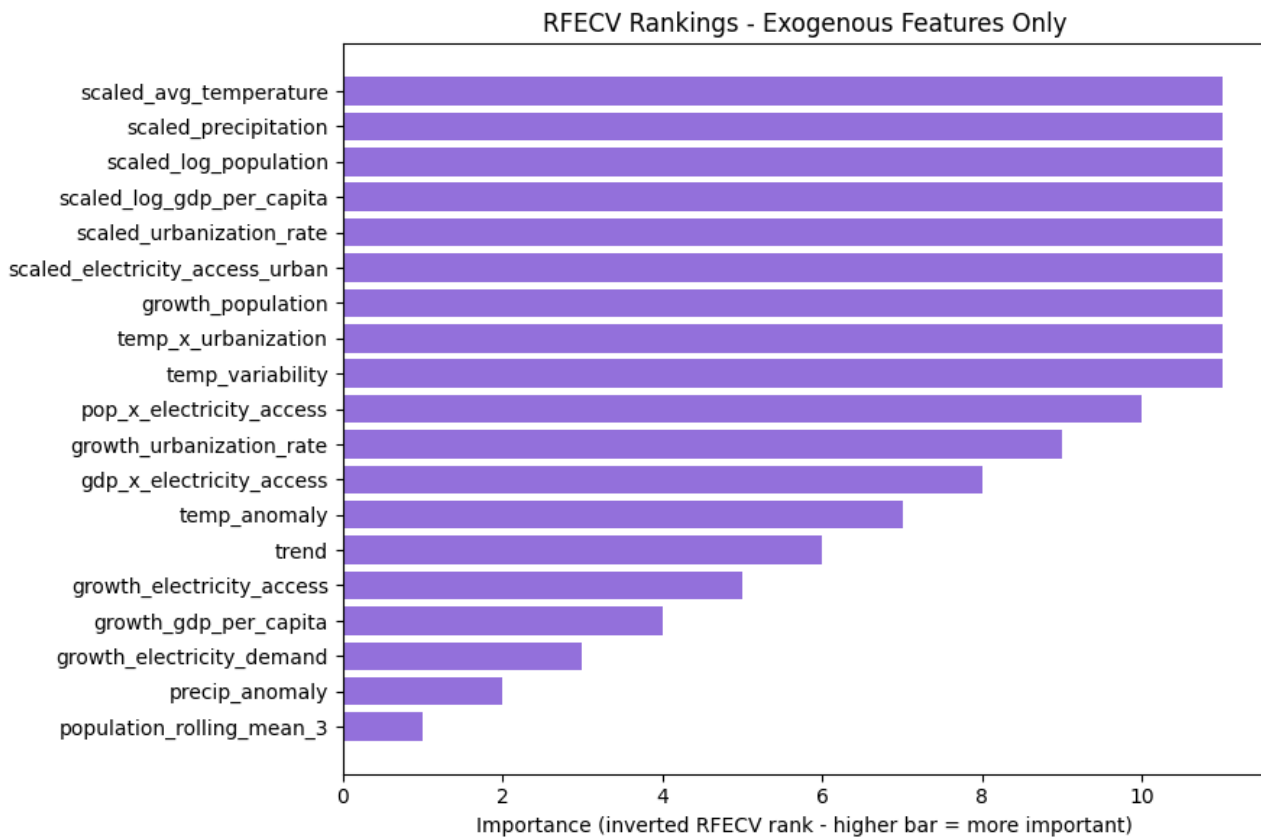


Figure 4.13: RFECV Feature Rankings - Exogenous Features Only (Higher Bar = More Important)

RFECV automatically selected 10 out of 19 exogenous features as optimal, with scaled log population ranked first, followed by scaled log GDP per capita, population growth rate, scaled urbanisation rate, and scaled average temperature.

With the lag features removed, the same two tree-based models were used again to rank the exogenous features. Figure 4.14 shows both models' scores side by side.

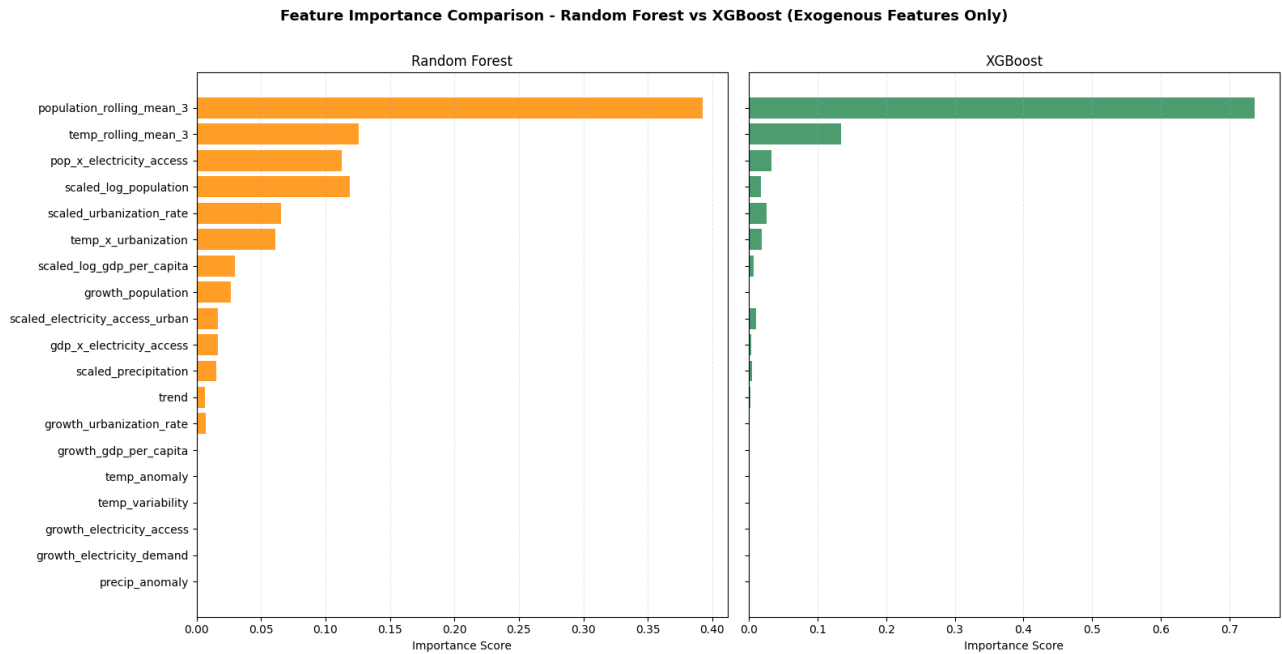


Figure 4.14: Feature Importance Comparison - Random Forest vs XGBoost (Exogenous Features Only)

This time, with the lag features out of the picture, the scores are more spread out and the differences between features are much easier to see. Both models agree clearly on the top two: the population rolling mean comes first in both (39.2% in Random Forest, 73.6% in XGBoost), and the temperature rolling mean comes second (12.6% in Random Forest, 13.5% in XGBoost). After that, Random Forest puts scaled log population third (11.9%) and the population-electricity access interaction fourth (11.2%), while XGBoost ranks population-electricity access third (3.3%) and scaled urbanisation rate fourth (2.6%).

At the bottom of both lists, growth rate features, anomaly features, and variability features all score very low. This tells us that year-to-year changes in things like GDP growth or temperature deviations are not very useful for predicting demand on their own. What matters much more is the overall level and trajectory of population and temperature.

Figure 4.15 shows the SHAP Feature Attribution plot for the Random Forest model trained on the exogenous features only.

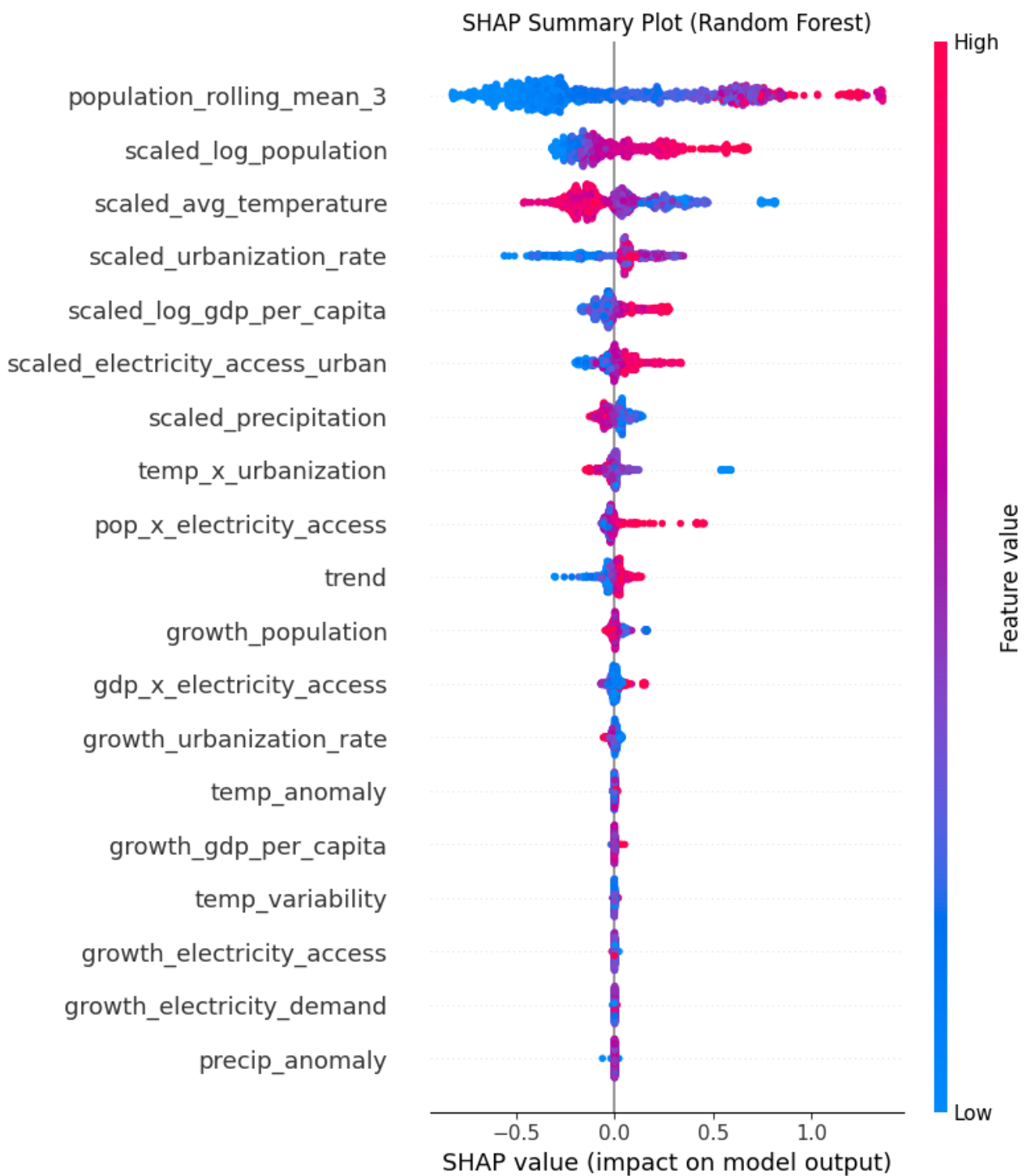


Figure 4.15: SHAP Feature Attribution Plot for the Random Forest Model on Exogenous Features Only

The exogenous SHAP plot told a more interesting and informative story than the full-feature version, because with the lag features removed the other features now had room to show their contributions.

Population rolling mean dominated the plot, with the widest spread of any feature. High values (red dots, large rolling population means) pushed predictions strongly to the right (upward), while small rolling means (blue) pushed them to the left (downward). This confirms that the medium-term population trajectory is the most important structural driver of electricity

demand among the exogenous variables.

Temperature rolling mean showed a clear negative direction. High temperature rolling mean values (red) sat on the left side of zero, meaning that countries or years with higher average temperatures were predicted to have lower demand. This is the opposite of what you might expect if you were thinking about cooling electricity demand. As explained in the previous chapter, the reason is that the hottest countries in Sub-Saharan Africa tend also to be the poorest and least electrified, so high temperature is acting as a proxy for low development in this dataset rather than as a direct cause of high cooling demand. This is an important contextual insight that SHAP makes clearly visible in a way that a simple correlation number could not convey.

Scaled log population showed a similar upward direction to the rolling mean, but with a narrower spread. The population-electricity access interaction spread across both positive and negative territory, which makes intuitive sense, a large population in a country with high electricity access drives demand up, but a large population in a country with very low electricity access shows a weaker predictive association with measured demand because many of those people are not connected to the grid.

Growth rate features, anomaly features, and variability features had dots clustered near zero with very little spread, confirming that they contributed little once the level-based features were already included.

4.6.3 Consolidated Exogenous Feature Ranking

Table 4.4 shows the consolidated ranking of all 19 exogenous features across the five selection methods.

Table 4.4: Consolidated Exogenous Feature Ranking (Normalised Scores, 0-1)

Feature	Pearson	MI	RFECV	RF	XGBoost	Combined
population_rolling_mean_3	0.975	0.897	0.000	1.000	1.000	0.774
scaled_log_population	1.000	1.000	1.000	0.303	0.024	0.665
temp_rolling_mean_3	0.352	0.703	1.000	0.320	0.184	0.512
pop_x_electricity_access	0.669	0.553	1.000	0.286	0.044	0.510
temp_x_urbanisation	0.526	0.514	1.000	0.154	0.026	0.444
scaled_urbanisation_rate	0.240	0.656	1.000	0.166	0.036	0.420
scaled_electricity_access_urban	0.569	0.369	1.000	0.041	0.015	0.399
scaled_log_gdp_per_capita	0.461	0.292	1.000	0.074	0.009	0.368
growth_population	0.132	0.443	1.000	0.067	0.001	0.329
scaled_precipitation	0.212	0.322	1.000	0.038	0.006	0.316
growth_urbanisation_rate	0.116	0.412	0.889	0.016	0.002	0.287
gdp_x_electricity_access	0.241	0.236	0.778	0.040	0.006	0.260
temp_variability	0.129	0.028	1.000	0.001	0.000	0.232
temp_anomaly	0.121	0.000	0.667	0.000	0.000	0.158
trend	0.202	0.000	0.556	0.015	0.000	0.155
growth_electricity_access	0.144	0.116	0.444	0.000	0.000	0.141
growth_electricity_demand	0.129	0.295	0.222	0.000	0.000	0.129
growth_gdp_per_capita	0.000	0.050	0.333	0.001	0.000	0.077
precip_anomaly	0.028	0.000	0.111	0.000	0.000	0.028

One point in this table is worth pointing out. The three-year rolling mean of population ended up with the highest combined score of 0.774, but it was actually the only feature that RFECV

rejected outright, giving it a score of 0.000 and ranking it last of all nineteen. It reached the top of the combined ranking purely because Random Forest and XGBoost both ranked it first by a wide margin. What this shows is that the rolling population mean carries strong information in a nonlinear form that the tree-based models can use, but adds little once the raw log population is already in a linear model, which is why RFECV dropped it. This is exactly the kind of disagreement that using several selection methods together was meant to expose.

The five features with the highest combined scores were selected for the exogenous feature set: population_rolling_mean_3 (0.774), scaled_log_population (0.665), temp_rolling_mean_3 (0.512), pop_x_electricity_access (0.510), and temp_x_urbanisation (0.444). Scaled urbanisation rate and urban electricity access both ranked just below the cutoff.

4.6.4 Final Feature Sets

Based on the two-stage feature selection process, three feature sets were defined for model training:

- **Model A (Lag-only):** lag_1_log_electricity_demand, lag_2_log_electricity_demand
- **Model B (Exogenous-only):** population_rolling_mean_3, scaled_log_population, temp_rolling_mean_3, pop_x_electricity_access, temp_x_urbanisation
- **Model C (Multivariate):** All seven features from Model A and Model B combined

4.7 Model Performance Results

Before the models could be trained, all rows with missing values in the target or in any of the seven selected features had to be dropped. Because the two-year lag feature needs two earlier years of demand data to exist, this removed the years 2000 and 2001 for every country. The final modelling dataset therefore contained 935 rows covering 47 countries over the period 2002 to 2021. Of these, 747 rows fell in the training and validation period and 188 rows fell in the 2018 to 2021 test period.

4.7.1 Hyperparameter Tuning Results

Table 4.5 shows the best hyperparameters identified through grid search for Random Forest, XGBoost, and LightGBM across each feature set.

Table 4.5: Best Hyperparameters from Grid Search Tuning

Model - Feature Set	Hyperparameter	Best Value
Random Forest - Model A	n_estimators, max_depth, min_samples_split	100, 15, 2
Random Forest - Model B	n_estimators, max_depth, min_samples_split	300, 10, 2
Random Forest - Model C	n_estimators, max_depth, min_samples_split	100, 15, 5
XGBoost - Model A	n_estimators, max_depth, learning_rate, reg_alpha, reg_lambda	300, 3, 0.05, 0, 1.0
XGBoost - Model B	n_estimators, max_depth, learning_rate, reg_alpha, reg_lambda	200, 6, 0.1, 0.1, 1.0
XGBoost - Model C	n_estimators, max_depth, learning_rate, reg_alpha, reg_lambda	300, 6, 0.1, 0, 2.0
LightGBM - Model A	n_estimators, max_depth, learning_rate, num_leaves	300, 6, 0.05, 31
LightGBM - Model B	n_estimators, max_depth, learning_rate, num_leaves	300, 9, 0.01, 15
LightGBM - Model C	n_estimators, max_depth, learning_rate, num_leaves	100, 6, 0.1, 31

4.7.2 Walk-Forward Validation Results

Table 4.6 presents the average performance metrics across all seven walk-forward validation iterations (predicting years 2011 to 2017) for all valid model and feature set combinations.

Table 4.6: Walk-Forward Validation Results – Average Performance Across 7 Iterations (2011-2017)

Feature Set	Model	MAE	RMSE	MAPE (%)	R ²
Model A: Lag-only	Linear Regression	0.0370	0.0540	6.42	0.9974
	XGBoost	0.0431	0.0626	5.20	0.9965
	Random Forest	0.0462	0.0659	5.61	0.9962
	Dynamic Fixed Effects Panel Regression	0.0694	0.0917	6.96	0.9924
	Fixed Effects Panel Regression	0.0694	0.0917	6.96	0.9924
	LightGBM	0.0908	0.2389	6.90	0.9494
Model B: Exogenous-only	Random Forest	0.1786	0.2650	31.72	0.9366
	XGBoost	0.1973	0.3164	30.76	0.9125
	LightGBM	0.2255	0.3011	45.97	0.9192
	Linear Regression	0.5526	0.6724	122.36	0.6095
	Fixed Effects Panel Regression	1.2660	1.4669	336.55	-0.8657
Model C: Full Multivariate	Linear Regression	0.0359	0.0530	5.73	0.9974
	Random Forest	0.0439	0.0666	5.20	0.9960
	XGBoost	0.0460	0.0671	5.80	0.9961
	LightGBM	0.0666	0.1050	9.35	0.9899
	Dynamic Fixed Effects Panel Regression	0.4490	0.4911	114.41	0.7700
	Fixed Effects Panel Regression	0.4490	0.4911	114.41	0.7700

Note: Dynamic Fixed Effects Panel Regression was not applied to Model B because this feature set contains no lag features, which are required for this model to be meaningful. LightGBM results are presented separately below due to its experimental status.

A few things stood out from the walk-forward results. In Model A, Linear Regression came out on top with an average RMSE of 0.0540 and an R² of 0.9974. This was an interesting result, because it suggested that the link between last year’s demand and this year’s demand is basically linear, and that heavier models like XGBoost and Random Forest were not adding anything useful on top of what a simple linear model could already do with the same two features.

In Model B the ranking flipped. Random Forest performed best here with an RMSE of 0.2650 and an R² of 0.9366, which suggests the climate and socioeconomic variables relate to demand in a nonlinear way that the tree-based model was able to pick up but Linear Regression could not. Linear Regression on Model B did poorly by comparison, with an RMSE of 0.6724 and a MAPE above 122%. The Fixed Effects Panel Regression did worst of all on Model B, producing a negative R² of -0.8657, which means it did worse than just predicting the average every time. This happened because with no lag variable in the model, the country-specific intercepts had nothing to anchor them to each country’s actual demand path.

In Model C, Linear Regression was again the strongest model, with an RMSE of 0.0530 and an R² of 0.9974. It is worth noting that Model C and Model A tied exactly on R² at 0.9974, and Model C was only very slightly ahead on RMSE and MAPE. So adding the five climate and socioeconomic features did not really improve accuracy in any meaningful way.

LightGBM underperformed on all three models compared to XGBoost and Random Forest. On Model A it recorded an average RMSE of 0.2389 and R² of 0.9494, substantially worse than the other models on the same feature set. Its performance improved on Models B and C where more features were available, but it still did not outperform the other tree-based models on any configuration. The reasons for this underperformance are discussed in Chapter 5.

4.7.3 Best Model Per Feature Set

Based on the walk-forward validation results, the best-performing model for each feature set was identified:

- **Model A:** Linear Regression (RMSE = 0.0540, $R^2 = 0.9974$)
- **Model B:** Random Forest (RMSE = 0.2650, $R^2 = 0.9366$)
- **Model C:** Linear Regression (RMSE = 0.0530, $R^2 = 0.9974$)

4.7.4 Final Test Set Results

Each of the three best models was retrained on the full training and validation period (2002 to 2017) and then evaluated once on the independent test set (2018 to 2021). Table 4.7 shows the final test set performance results.

Table 4.7: Final Test Set Performance (Independent Test Period 2018–2021)

Feature Set	Best Model	MAE	RMSE	MAPE (%)	R^2
Model A: Lag-only	Linear Regression	0.0441	0.0678	6.23	0.9963
Model B: Exogenous-only	Random Forest	0.2416	0.3356	37.89	0.9085
Model C: Full Multivariate	Linear Regression	0.0436	0.0680	6.01	0.9962

All three models generalised well to the unseen test period. The test set results were consistent with the walk-forward validation results in terms of relative model rankings, which confirmed that the walk-forward validation was not overfitting. the Lag-only Model (Model A) and the Multivariate Model (Model C) produced very similar test set results, with RMSE values of 0.0678 and 0.0680 respectively, and R^2 values of 0.9963 and 0.9962 respectively. Model B (exogenous-only) performed considerably weaker with RMSE of 0.336 and log-scale MAPE of 37.89%. This confirms that within these 47 countries separated by time, climate and socioeconomic variables on their own produced less precise year-by-year predictions than models that included historical demand. It should be noted that this evaluation was carried out on the same set of countries used for training, separated only by time period. There was no out-of-country validation, so the results should be interpreted as evidence of structural predictive association within the observed countries and time period, not as evidence of generalisation to unseen countries or contexts.

4.7.5 Three-Way Feature Set Comparison

Figure 4.16 shows the three-way comparison of RMSE and R^2 across the three feature sets on the test period.

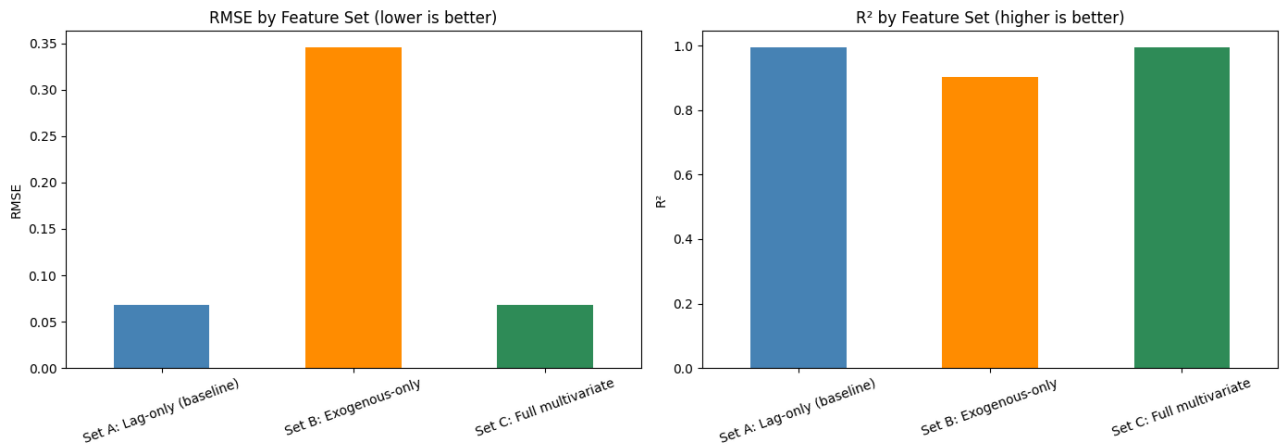


Figure 4.16: Three-Way Feature Set Comparison on Independent Test Set (2018-2021)

The comparison charts made clear that the Exogenous-only Model (Model B) had a substantially higher RMSE and lower R^2 than Models A and C, while Models A and C were almost indistinguishable from each other in terms of forecasting accuracy. The RMSE improvement from Model A to Model C was -0.36% , meaning that adding the five climate and socioeconomic features to the lag-only model produced essentially no change in overall forecast accuracy. This was a significant finding and is discussed in depth in Chapter 5.

4.8 Permutation Importance Results

After the final model was selected and trained on the complete dataset (2000 to 2021), permutation importance analysis was carried out on the test period (2018 to 2021) to validate which features the model was genuinely relying on for its predictions. Table 4.8 shows the permutation importance results.

Table 4.8: Permutation Importance Results - Final Model on Test Set (2018-2021)

Feature	Mean R^2 Drop	Std Dev
lag_1_log_electricity_demand	2.3296	0.1619
lag_2_log_electricity_demand	0.0156	0.0015
scaled_log_population	0.0003	0.0001
pop_x_electricity_access	0.0000	0.0000
population_rolling_mean_3	0.0000	0.0000
scaled_avg_temperature	0.0000	0.0000
temp_x_urbanisation	-0.0000	0.0000

Figure 4.17 shows the permutation importance results as a bar chart with error bars representing the standard deviation across 30 repetitions.

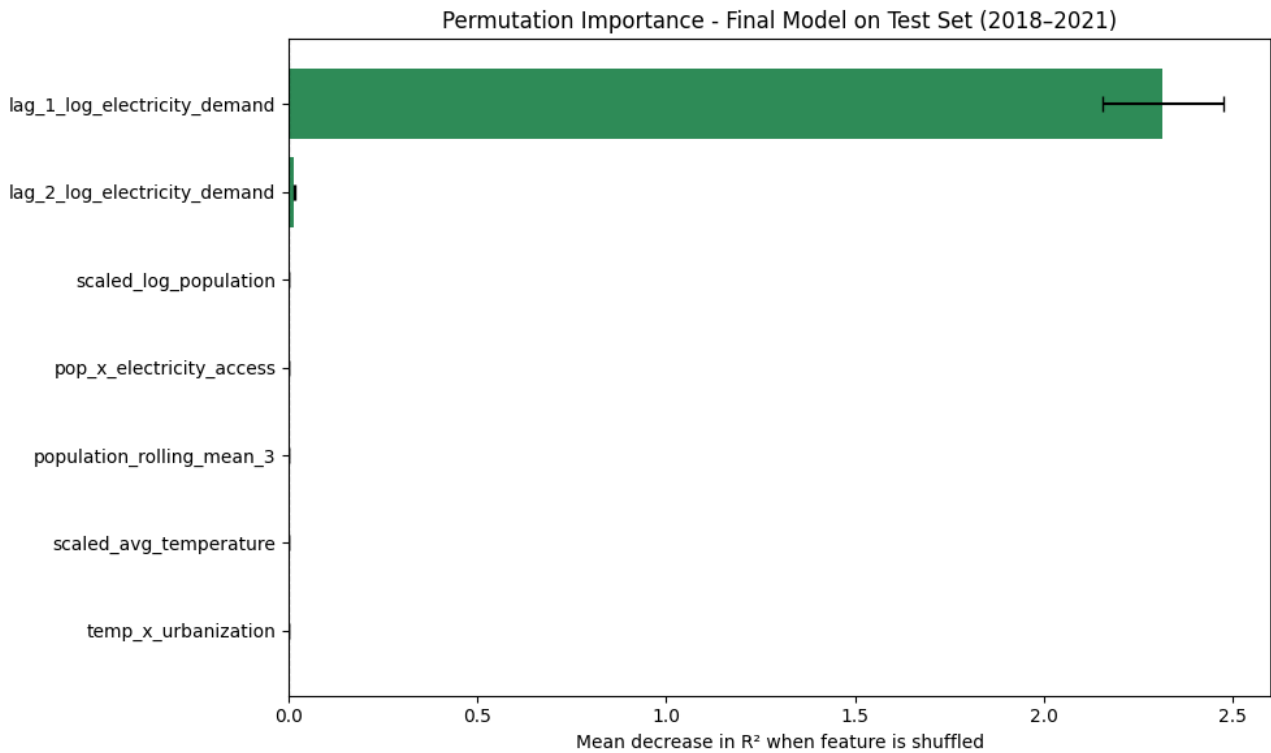


Figure 4.17: Permutation Importance of Features in the Final Model - Test Set (2018-2021)

The permutation importance results confirmed the dominance of `lag_1_log_electricity_demand` in the final model. Shuffling this feature caused an average drop in R^2 of 2.316, with a small standard deviation of 0.161 across 30 repetitions, confirming that this result was stable and not sensitive to the particular random shuffles used. Shuffling `lag_2` caused a much smaller but still positive drop of 0.014, confirming that the two-year lag added some marginal value. All five climate and socioeconomic features produced R^2 drops at or very close to zero, confirming that the final model was deriving almost all of its predictive power from the lag features rather than from the structural variables.

4.9 Final Model Selection

Based on the walk-forward validation results, the test set performance, and the permutation importance analysis, Linear Regression on Model C (the full multivariate feature set) was selected as the final model. Table 4.9 summarises the test set performance of the key competing models.

Table 4.9: Summary of Key Model Performances on Test Set - Basis for Final Selection

Feature Set	Model	MAE	RMSE	MAPE (%)	R^2
Model A: Lag-only	Linear Regression	0.0441	0.0678	6.23	0.9963
Model B: Exogenous-only	Random Forest	0.2416	0.3356	37.89	0.9085
Model C: Full Multivariate	Linear Regression	0.0436	0.0680	6.01	0.9962
Model A: Lag-only	XGBoost	0.0494	0.0763	5.85	0.9949
Model A: Lag-only	Random Forest	0.0503	0.0798	5.97	0.9942

Linear Regression on Model C was selected because it achieved the best overall balance of forecasting accuracy, interpretability, and alignment with the study's research objectives. It

produced a log-scale MAPE of 6.01% and a very high R^2 of 0.9962 on the test set. While its RMSE (0.0680) was marginally higher than Model A's Linear Regression (0.0678) by just 0.0002, the difference was negligible and did not justify choosing the lag-only model over the full multivariate model. Model C was preferred because it included climate and socioeconomic variables that allowed the study to address the core research questions about what drives electricity demand, and it provided more informative model coefficients and feature importance scores for decision-makers to act on.

The final model was retrained on the complete 2000-2021 dataset, saved using `joblib`, and deployed in the Streamlit decision-support dashboard.

4.10 Decision-Support Dashboard

The Streamlit dashboard was successfully developed and deployed as a standalone application. It loaded the pre-trained final model, the saved feature list, and the engineered dataset automatically at startup. Figures 4.18 to 4.25 show screenshots from the four main pages of the dashboard.



Figure 4.18: Dashboard Overview Page - Electricity Demand Trends

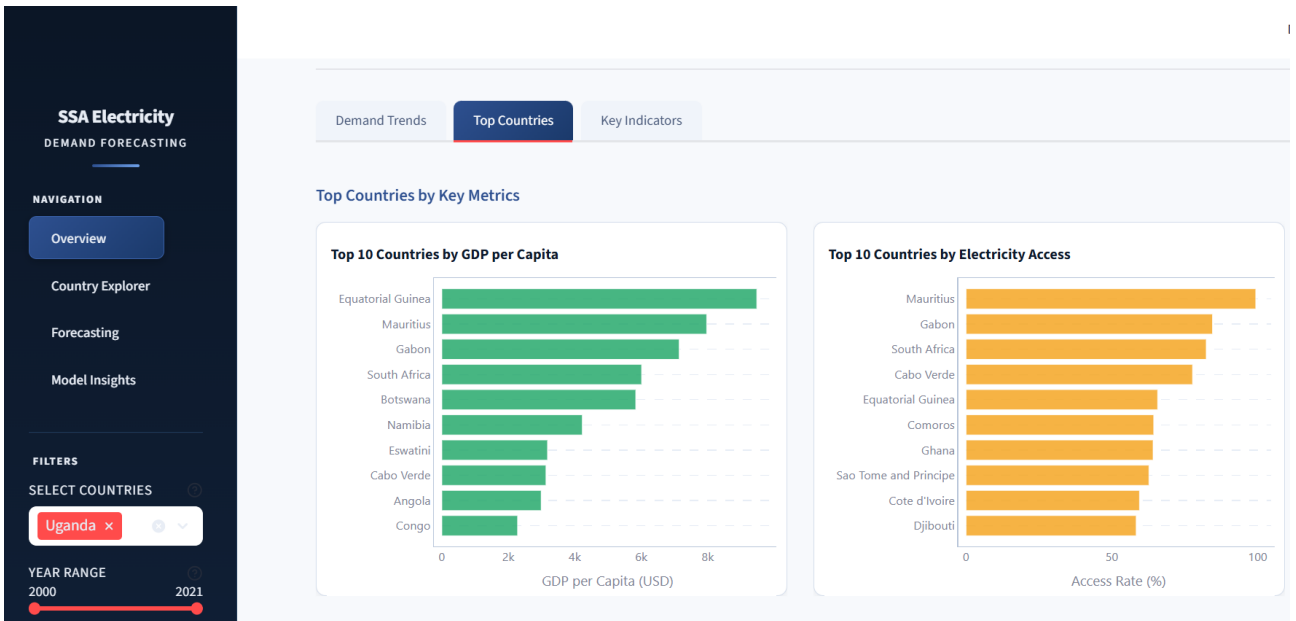


Figure 4.19: Dashboard Overview Page - Electricity Top Countries

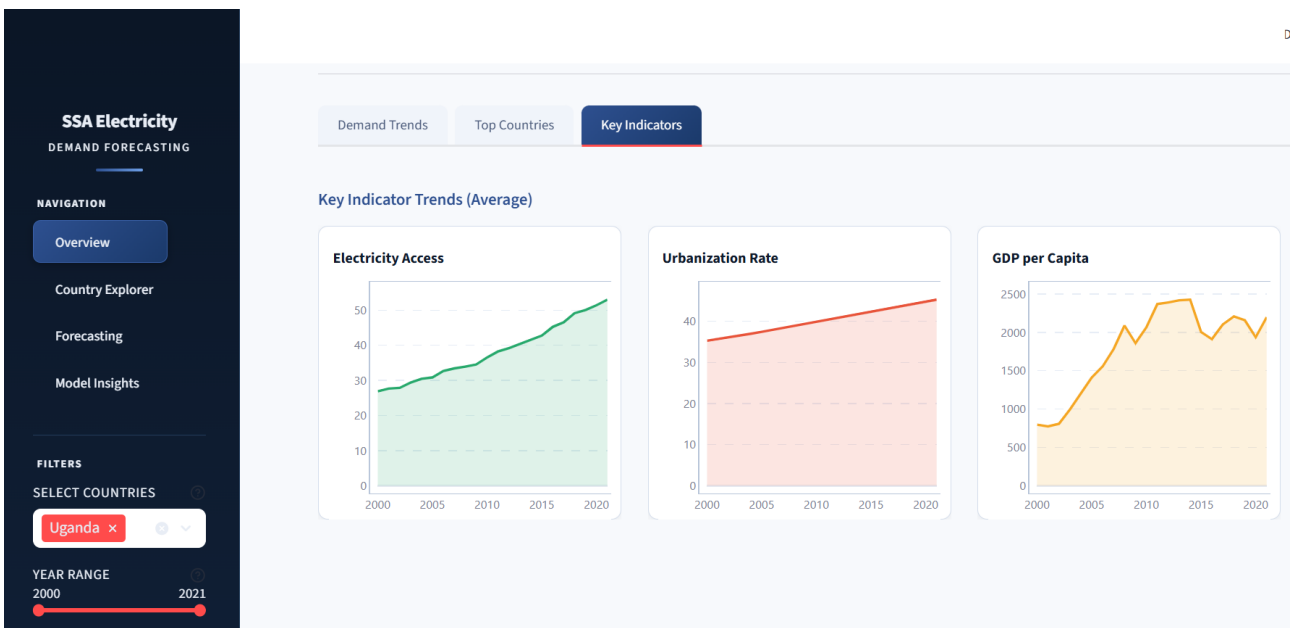


Figure 4.20: Dashboard Overview Page - Electricity Key Indicators

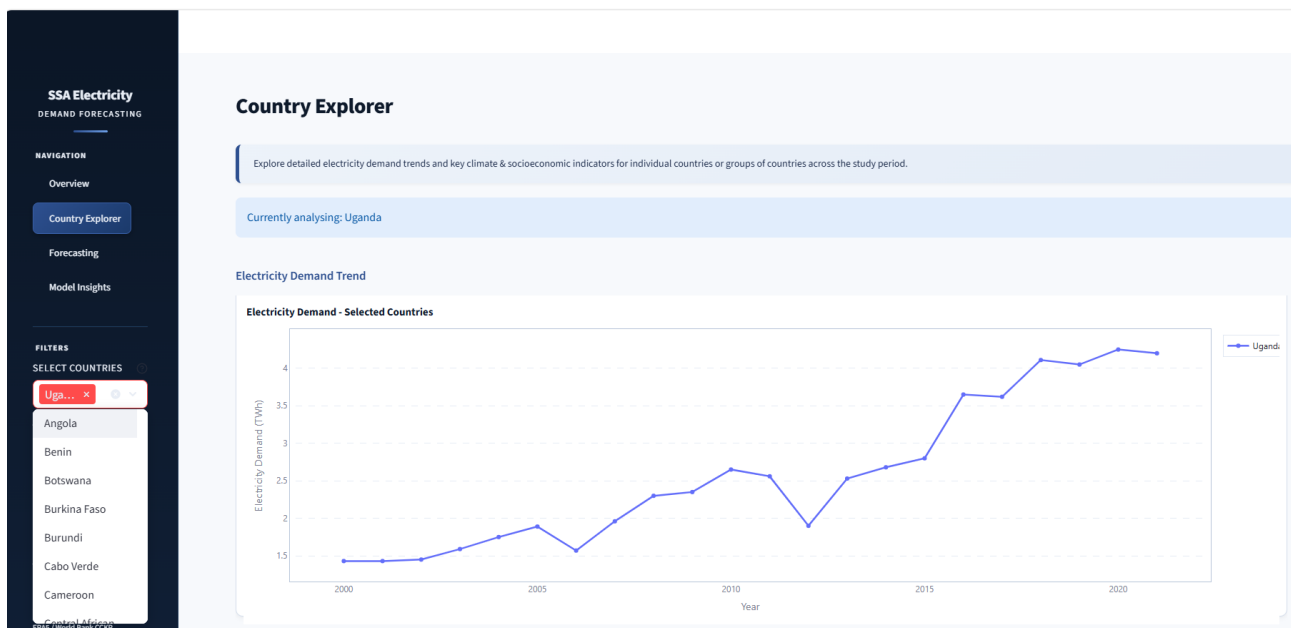


Figure 4.21: Dashboard Country Explorer Page - Country-Level Demand (Uganda as the selected country)

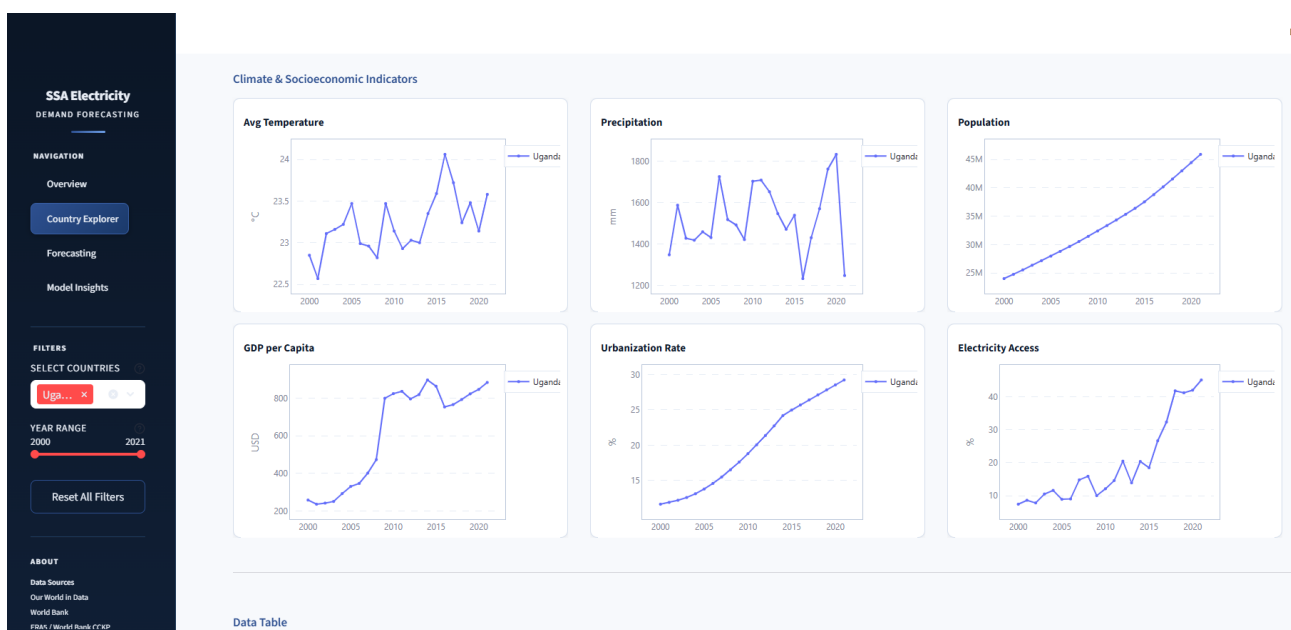


Figure 4.22: Dashboard Country Explorer Page - Country-Level Indicator Trends (Uganda as the selected country)



Figure 4.23: Dashboard Forecasting Page - Actual vs Predicted Demand for a Selected Country

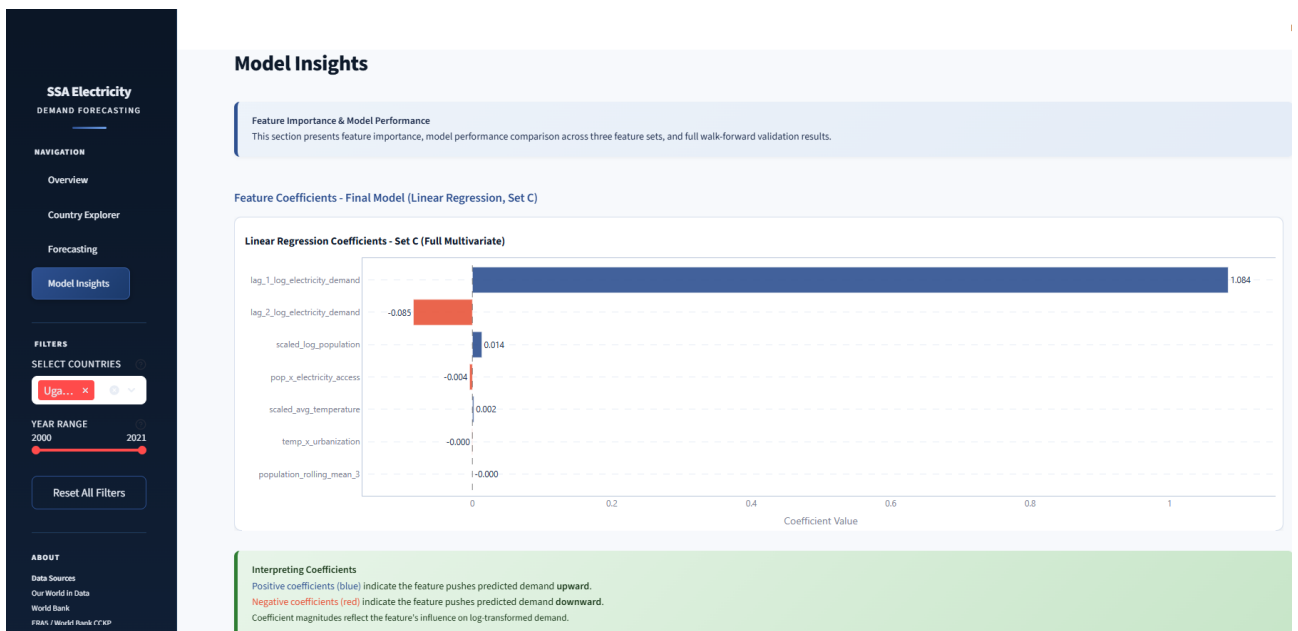


Figure 4.24: Dashboard Model Insights Page - Feature Coefficients

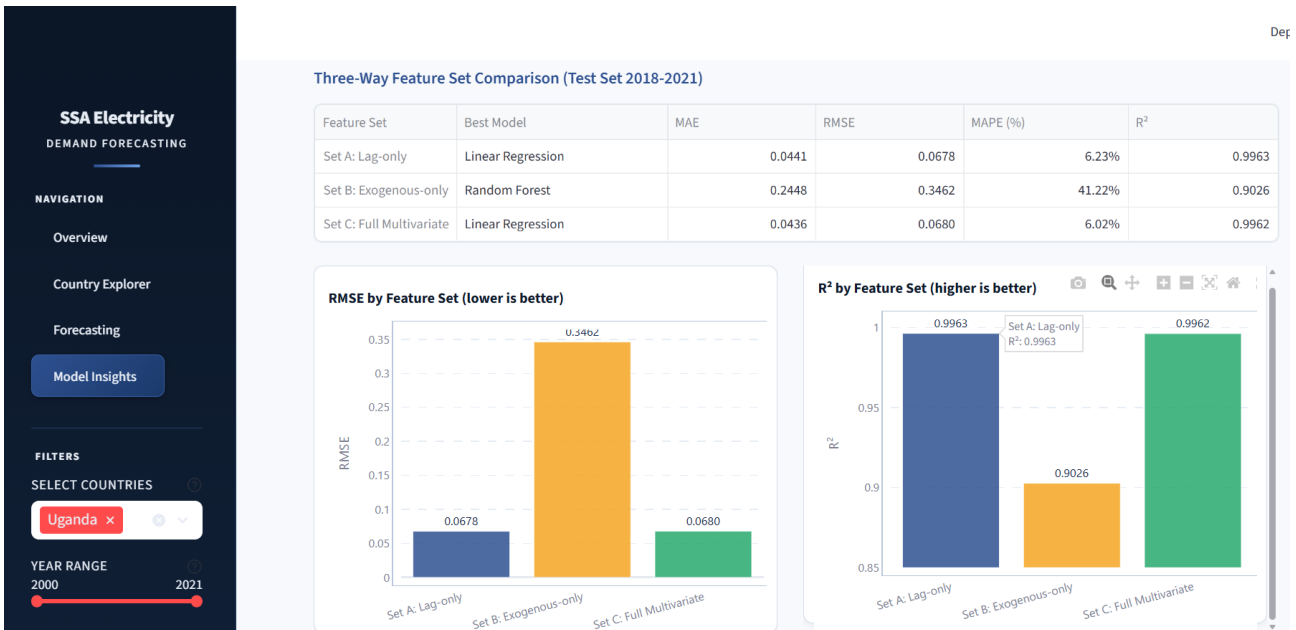


Figure 4.25: Dashboard Model Insights Page - Model Performance Comparison

The dashboard was built as an exploration and transparency tool rather than a forward-forecasting one. It presents the historical data, the model's back-tested predictions over the 2018 to 2021 test period, and the model's internal workings, but it does not generate forecasts beyond 2021, since that would require future values of population and temperature that this study did not project. The dashboard provided users with four main sections. The Overview page showed the total electricity demand trend across all 47 countries, the top 10 countries by average demand, and pooled trends in electricity access, urbanisation, and GDP per capita. The Country Explorer page allowed users to select any combination of countries and a year range and explore their demand trends alongside six climate and socioeconomic indicator charts and a downloadable data table. The Forecasting page displayed actual versus predicted electricity demand for any selected country over the full study period, with the test period predictions highlighted, and showed forecast accuracy metrics for that country. The Model Insights page presented the final model's feature coefficients, the three-way feature set comparison results, and the full walk-forward validation summary table for all 17 valid model-feature set combinations.

Chapter 5

Discussion

5.1 Introduction

This chapter interprets and discusses the results presented in Chapter 4. Rather than simply restating the numbers, the aim here is to explain what those results mean, why they turned out the way they did, and what they say about electricity demand in Sub-Saharan Africa more broadly. The discussion is organised around the main findings of the study, starting with the most striking result, the dominance of lag features, before moving on to the role of individual structural drivers, the performance of the different model types, and what the overall findings mean for energy planning and policy in the region. The chapter also compares the findings with the existing literature and acknowledges the limitations of the study.

5.2 Why Lag Features Dominate Annual Electricity Demand Forecasting

The clearest and most consistent finding across all analytical stages was the near-total dominance of the one-year lag feature in predicting electricity demand. The one-year lag of log-transformed electricity demand had a Pearson correlation of 0.999 with the target variable, and the permutation importance analysis showed that shuffling this single feature caused an average drop in R^2 of 2.316, a massive change that dwarfed the contribution of all other features combined. The two-year lag contributed a further, much smaller drop of 0.014. In contrast, shuffling any of the five climate and socioeconomic features caused drops at or near zero.

This result is not surprising when you think about what annual electricity demand actually measures. The total amount of electricity consumed in a country in a given year reflects the accumulated effect of that country's entire energy infrastructure, industrial base, household appliance ownership, urbanisation level, and electrification coverage at that point in time. None of these things change dramatically from one year to the next. A country with 10 million electricity customers in 2015 will very likely have 10.3 or 10.5 million in 2016, not 20 million. The electricity system, the economy, and the population all evolve slowly and continuously. This means that knowing how much electricity a country consumed last year gives you almost all the information you need to predict how much it will consume this year, simply because almost everything that determines demand stays roughly the same from one year to the next [13, 31].

This phenomenon is well documented in the energy forecasting literature. [4] described electricity demand as exhibiting strong autoregressive behaviour across all time horizons, meaning

that past demand is always a powerful predictor of future demand. [12] similarly noted that lag features consistently outperform other predictors in annual energy forecasting contexts, particularly when the underlying socioeconomic and climate conditions are relatively stable. What this study adds is a precise quantification of just how dominant this effect is in the Sub-Saharan African context, the one-year lag alone accounted for nearly all of the model’s predictive power, leaving almost nothing for the structural variables to explain on top of it.

It is important to be clear about what this does and does not mean. It does not mean that population, GDP, temperature, and electricity access are unimportant for electricity demand. It means that once you know last year’s demand, those variables do not tell you much more about this year’s demand than you already know. This is because the effects of population growth, economic development, and electrification expansion are already embedded in the historical demand trajectory itself, they are what is associated with last year’s demand levels, and their continued influence will show up in next year’s demand through the same channel.

This finding also has an important methodological implication. The lag dominance result is specific to the annual panel forecasting setting used in this study. At shorter time horizons (monthly or daily), the autoregressive signal is still present but does not absorb the exogenous variables so completely, because weekly or seasonal climate variation and economic shocks produce changes that are not yet captured in the most recent annual observation. Several of the studies reviewed in Chapter 2, such as [10] and [7], worked at higher frequencies and did find meaningful contributions from climate and socioeconomic variables even after controlling for lagged demand. This study’s result should therefore not be read as a general finding that structural variables are unimportant but rather it is a finding that is specific to annual data in a slowly-evolving economic environment.

There is also something that this study cannot conclude. Because the lag dominance was so complete, it is not possible to say from this study alone how much of the demand variation is truly caused by historical persistence versus how much is caused by the slow-moving structural variables that are correlated with lagged demand. This is a known limitation of autoregressive models in panel data, the lag feature absorbs the cumulative effect of all correlated predictors, making it hard to separate their individual contributions [18]. A different research design, such as an interrupted time series or a natural experiment, would be needed to isolate these effects. The practical consequence is that a planning unit with no modelling capacity at all can produce a competent one-year-ahead national forecast from last year’s figure and a growth rate. The value of a fuller model lies elsewhere, in longer horizons and in scenario work.

5.3 Why Population Is the Most Important Structural Driver

When the lag features were removed and the models were evaluated on climate and socioeconomic variables only (Model B), population-related features emerged as the strongest structural predictors of electricity demand. The three-year rolling mean of population (`population_rolling_mean_3`) had the highest combined selection score of all exogenous features (0.774), followed by the raw annual log-population (0.665). The Random Forest model assigned 39.2% of its total importance to `population_rolling_mean_3`, and XGBoost assigned 73.6% to the same feature.

The reason population is so important is straightforward. More people means more households, more businesses, more schools, and more hospitals, all of which need electricity. In Sub-Saharan Africa, where the population was growing at among the fastest rates in the world during the

study period, this relationship was particularly strong. Countries like Nigeria, Ethiopia, and Tanzania, which all had fast-growing populations during 2000 to 2021, also showed consistent upward trends in electricity demand, even though their per-capita consumption was relatively low [2].

What is more interesting is that the three-year rolling mean of population consistently outperformed the raw annual population value across all selection methods. This tells us something meaningful about how population growth affects electricity demand. The total population in a given year is one snapshot, but electricity infrastructure i.e power lines, substations, household connections, takes time to build and expand. The three-year rolling mean better captures the medium-term trajectory of population growth that electricity systems have had time to respond to, rather than the single most recent year's figure, which may be noisy or may not yet have translated fully into new demand connections. This finding is consistent with the idea that the electricity system lags behind population growth and adjusts gradually rather than instantly [11].

Population was also important through the interaction term with electricity access (`pop_x_electricity_access`). A large population only translates into high electricity demand if those people are actually connected to the electricity grid. In countries where electrification was low and population was high, the interaction term captured the fact that the relationship between population and demand was weaker than in countries with high access rates. This nuance would have been missed if only raw population and access rate had been used as separate features.

The finding that the rolling mean of population is more predictive than the raw annual figure has not been widely reported in the existing electricity demand forecasting literature, which typically uses current-year population as a predictor [11, 23]. This result suggests that the infrastructure lag hypothesis has empirical support in the Sub-Saharan African context: grid expansion responds to population trends over multiple years, not to single-year changes. This has a practical implication for energy planners. If demand responds to a smoothed population signal rather than an instantaneous one, then demographic projections made three years in advance may actually be more useful inputs for infrastructure planning than the most recent single-year population estimate.

It should be noted, however, that the finding of population being the strongest predictor in Model B is a predictive association, not a causal relationship. Population size is highly correlated with GDP, urbanisation, and electrification level across Sub-Saharan African countries. The population feature may therefore be acting partly as a proxy for these other structural factors that were not independently selected due to multicollinearity. This is consistent with the observation that scaled log population was the second-ranked feature, very close in score to population rolling mean, and that both carry similar information. For planners this points somewhere specific: national demand projections should be built on the population projections the statistics office already produces, and should be revised when those projections are revised.

5.4 Why GDP Contributes Less Once Lags Are Included

GDP per capita was not selected as one of the top five exogenous features, with a combined score of 0.381 that fell just below the top-five threshold. This was somewhat unexpected given that GDP is frequently cited in the electricity demand forecasting literature as one of the most important predictors of energy consumption [22, 23].

Two main reasons explain why GDP performed less strongly in this study. The first is multicollinearity. GDP per capita was strongly correlated with electricity access rate (Pearson $r = 0.77$) in this dataset, meaning that wealthier countries also tended to have higher electricity access, and vice versa. When both variables were present in the feature selection process, the selection methods tended to favour whichever one of the pair had a slightly stronger relationship with demand and discard the other. In this case, electricity access-related features came out slightly stronger than GDP in most methods, so GDP was effectively crowded out.

The second reason is more fundamental and relates to the lag dominance finding. When the full multivariate model (Model C) was trained, the lag features absorbed most of the predictive signal that would otherwise have come from GDP. This is because GDP growth tends to be highly autocorrelated, meaning that a country's GDP this year is very closely related to its GDP last year. If last year's electricity demand already reflects last year's GDP level, and last year's GDP is itself closely related to this year's GDP, then including this year's GDP in the model provides very little additional information beyond what the lag feature already captured. In other words, the lag feature acts as a compressed summary of all the structural variables, including GDP, and once the lag is in the model there is not much left for GDP to explain [4, 12].

This does not mean that GDP is unimportant for electricity demand. The Model B results, where the lag features were excluded, showed that GDP per capita was a relevant contributor to the exogenous model. Studies that have used structural models without lagged demand as a reference, such as [22] and [23], consistently found strong positive GDP-demand relationships. The difference in this study is the methodological choice to include lagged demand as a feature. This choice, which is standard in dynamic forecasting models, effectively transforms the research question from "what level of demand is associated with current GDP?" to "how much additional information does current GDP provide once we already know last year's demand?" The answer, in this dataset, is very little. This is not a contradiction of the existing literature, it is a clarification of what the existing literature found, applied to a different modelling context.

One important implication of this is that the coefficients produced by Model C on the GDP variable should not be interpreted as an estimate of the GDP-demand elasticity. In an autoregressive model, the coefficient on a structural predictor reflects its marginal contribution after the lag has already accounted for most of the variation, not its total relationship with demand. Researchers who need an elasticity estimate should use a model that does not include the lagged dependent variable, such as a static Fixed Effects Panel model fitted on the same data.

5.5 Why Temperature Shows a Negative Relationship with Demand

One of the more counterintuitive findings of this study was that average temperature showed a negative correlation with electricity demand ($r = -0.213$ for Pearson and $r = -0.229$ for Spearman). This means that hotter countries, on average, tended to have lower electricity demand. This is the opposite of what might be expected based on studies from temperate regions, where higher temperatures are strongly associated with increased cooling-related electricity consumption [10, 29].

The explanation for this finding lies in the structural characteristics of Sub-Saharan Africa. The countries with the highest average temperatures in the dataset, countries in the Sahel and the Horn of Africa such as Mali, Niger, Somalia, and Djibouti, are also among the poorest and least

electrified countries in the region. They have very low electricity access rates, limited industrial activity, and small formal economies. As a result, their total electricity demand is very low, not because high temperatures reduce demand, but because high temperatures in this context are a proxy for a broader set of conditions associated with low electricity consumption.

The countries with the highest electricity demand, by contrast (South Africa, Nigeria, and Mozambique) are at lower or more moderate average temperatures and have much more developed electricity infrastructure. South Africa in particular, with its large mining sector and relatively temperate climate, is a major outlier that pulls the cross-country relationship between temperature and demand in a negative direction.

This finding highlights an important point about applying temperature as a predictor in cross-country panel forecasting. In a within-country time series context, higher-than-usual temperatures in a given year might well be associated with higher cooling demand. But in a cross-country comparison, the between-country temperature-demand relationship is confounded by the much larger differences in economic development and electrification levels. The temperature anomaly feature (which captured how much a given year’s temperature deviated from each country’s own long-run average) was designed to capture the within-country effect, though its contribution in this study was also small. Future research could more carefully separate these within-country and between-country effects using country-specific interaction terms or by focusing analyses within subgroups of countries with similar development levels [31].

This result directly addresses Research Question 1 in an unexpected way. The question asked how engineered climate features influenced forecasting. The answer, in this study, is that the raw temperature signal was useful primarily as a cross-country structural correlate of development level rather than as a direct climate influence on energy use. The temperature rolling mean performed better than the raw annual temperature, likely because it better captures the stable, slowly-changing thermal environment of each country rather than picking up on year-to-year noise. The temperature-urbanisation interaction was also selected, reflecting the idea that temperature’s relationship with demand is conditional on how urbanised the country is. Together, these engineered versions of the temperature variable were more informative than the raw annual value, which is itself a contribution to the feature engineering literature for SSA electricity forecasting. This sign should be expected to flip as incomes rise and cooling becomes common, so the relationship found here should be treated as a description of the study period rather than a stable parameter to carry forward.

5.6 What the Exogenous-Only Results Tell Us About Structural Demand Drivers

Model B, the exogenous-only model, produced a test set R^2 of 0.909 with Random Forest as the best-performing model. On one hand, an R^2 of 0.909 is quite strong in many forecasting settings. On the other hand, the MAPE of 37.89% was still very high, suggesting that while the model correctly captured much of the broad cross-country and long-term variation in demand, it was not accurate enough for practical forecasting use.

What this result tells us is that the climate and socioeconomic variables do carry substantial predictive information about electricity demand in Sub-Saharan Africa, but not enough to produce precise year-by-year forecasts on their own. This distinction matters: an R^2 of 0.909 means that about 91% of the variation in log-transformed demand across countries and years can be accounted for by the five exogenous features selected. But a log-scale MAPE of 37.89% means that the model’s predictions are frequently off by large amounts at the country level.

These two numbers are not contradictory: the model can capture the broad cross-country pattern (rich, urban, well-connected countries have higher demand) while still being a poor predictor of the exact year-by-year value for any specific country. This is a classic distinction in panel data modelling between between-entity fit and within-entity fit. Model B appears to have strong between-country fit but weak within-country fit, which is exactly what you would expect when the exogenous variables are more stable across time than across countries. The structural variables are good at explaining why one country has higher demand than another (South Africa has more demand than Niger because it has a larger economy, a cooler climate, and better electricity infrastructure), but they are less good at explaining exactly how much any given country’s demand will change from one year to the next, because year-to-year changes are dominated by the autoregressive persistence that only the lag features can capture [4, 7].

This distinction between explaining structural differences and predicting year-to-year changes is important for how the results of this study should be used. The exogenous-only model (Model B) is most useful for understanding the structural associations between climate and socioeconomic variables and electricity demand within the observed countries, or for situations where no historical demand data is available. The model was trained and tested on the same 47 countries separated only by time, so claims about its performance for new or unseen countries should be made with caution [12].

The Fixed Effects Panel Regression performed particularly poorly on Model B, producing a negative R^2 of -0.8657 . This is because the Fixed Effects model estimates a separate intercept for each country to account for permanent country-specific differences. When lagged demand is not included as a variable, the model has to predict each country’s level of demand using only the current year’s exogenous variables, and the country fixed effects absorb so much of the cross-country variation that the within-country variation (which is what the exogenous variables are supposed to explain) becomes very hard to fit. This is a known limitation of Fixed Effects models in short panels where the cross-sectional variation is large relative to the time-series variation [18].

5.7 Why Linear Regression Outperformed More Complex Models

One of the findings was that Linear Regression outperformed Random Forest, XGBoost, and LightGBM on both Model A and Model C. On Model A, Linear Regression achieved a walk-forward RMSE of 0.0540, against 0.0626 for XGBoost and 0.0659 for Random Forest. On Model C, Linear Regression reached 0.0530, against 0.0677 for XGBoost and 0.0669 for Random Forest.

The reason for this is connected to the lag dominance finding. The relationship between last year’s electricity demand and this year’s demand is, at the annual level, essentially a linear relationship. A country that consumed 10 TWh last year will, with high probability, consume something close to 10.3 or 10.5 TWh this year. This relationship is stable, consistent, and approximately linear. Linear Regression is perfectly designed to capture this kind of relationship, and it does so without introducing any additional complexity. More complex models like Random Forest and XGBoost, which are designed to find nonlinear patterns and interactions in data, introduce additional variance when applied to a fundamentally linear problem, which is why they performed slightly worse [24, 25].

On the Exogenous-only Model (Model B), the picture was reversed. Random Forest outperformed Linear Regression by a wide margin (RMSE of 0.2650 against 0.6724). This makes

sense because the relationship between structural variables like population, temperature, and GDP and electricity demand is not linear. A doubling of population does not simply double demand; the relationship between urbanisation and demand depends on the level of electricity access; and the effect of temperature on demand depends on whether the country is hot or cold to begin with. Tree-based models are much better at capturing these kinds of conditional, nonlinear relationships, which explains why they performed better when the lag features were not present [12, 10].

LightGBM underperformed on all three feature sets. Its worst result was on Model A, where it recorded an average RMSE of 0.2389 and an R^2 of 0.9494, dramatically worse than the other models on the same feature set. The most likely explanation is that LightGBM’s leaf-wise tree growing strategy, which builds trees by splitting the leaf with the largest loss rather than growing trees level by level, creates deeper and more irregular trees that are more prone to overfitting when only two features are available. With just lag_1 and lag_2 as inputs, LightGBM found complex, high-variance tree structures that did not generalise well to the validation set. On Model B and Model C, its performance was more competitive but still did not surpass XGBoost or Random Forest, suggesting that the additional complexity of LightGBM’s approach did not offer meaningful benefits for this particular dataset [15].

5.8 Why Model C Was Selected Despite Near-Identical Performance to Model A

The comparison between the Lag-only Model (Model A) and the Multivariate Model (Model C) produced one of the most important and nuanced findings of this study. The difference in test set RMSE between the two was just 0.0002 (0.0678 for Model A versus 0.0680 for Model C), a difference of less than 0.5%. Statistically, this is negligible. Yet Model C (the full multivariate model) was selected as the final model rather than the simpler Model A.

This decision was made on three grounds that go beyond the raw accuracy metrics. First, the two models answer different questions, and only one of them can answer the question that matters most for planning. Model A predicts next year’s demand from last year’s. That works only where last year’s demand is known, which restricts it to a one-year horizon on countries with current data. Energy planning does not operate on a one-year horizon. A generation plant commissioned today will be built against a demand estimate for the 2030s, and for any year beyond the next one, the lag input simply does not exist. A lag-only model faced with that horizon has to feed its own predictions back in as inputs, and the error compounds at every step. Model C’s structural variables are the only inputs in this study that can be projected forward independently, because national statistics offices already publish population projections and climate models already publish temperature projections. Nobody publishes a projection of next decade’s electricity demand, which is the thing being asked for [1, 2].

Second, the climate and socioeconomic variables in Model C provide information that the lag features cannot. The lag features tell the model how much demand there was last year, but they do not tell it why or what structural factors are driving the trend. The exogenous variables (population trends, temperature, electricity access interactions) provide that context. This contextual information is what makes the model useful for decision-makers, who need to understand not just what demand is likely to be, but what is driving it and how changing circumstances might affect it [34].

Third, the negligible performance difference between Model A and Model C was itself a meaningful result. It confirmed that adding carefully selected climate and socioeconomic features

to a lag-based model did not meaningfully degrade performance, with RMSE moving by 0.3 percent, which means the full multivariate model can be recommended to energy planners without any concern that the additional variables are introducing noise or reducing forecast reliability. This is not always the case, adding irrelevant variables to a forecasting model can sometimes worsen performance, but the careful two-stage feature selection process used in this study ensured that only genuinely informative exogenous features were included in Model C [12, 19].

5.9 What the Climate and Socioeconomic Variables Contribute

The comparison between Model A and Model C raises an obvious question. If adding five climate and socioeconomic variables changes test set RMSE from 0.0678 to 0.0680, and if permutation importance shows all five contributing approximately zero, in what sense do these variables matter at all? The question deserves a direct answer rather than a defence of the study design.

The answer is that predictive contribution and structural contribution are different things, and this study measured both.

On predictive contribution, the finding is negative and is reported as such. For one-year-ahead forecasting, with last year's demand available, the five exogenous variables add nothing measurable. The RMSE difference is 0.3 per cent in the wrong direction, and permutation importance on the unseen test period returned drops at or near zero for all five, with one slightly negative. This study does not claim otherwise, and any reader who takes only this result away has taken away something true.

On structural contribution, the finding is positive and it comes from Model B. Five variables carrying no information whatsoever about past demand accounted for approximately 91 per cent of the variation in log demand across 47 countries and 20 years. That is not a small number and it is not an artefact. It means that if all demand history for a country were erased, these five variables would still place that country's demand level roughly correctly. Population size, how many people are connected, how urban the country is and how warm it is: between them, these largely determine how much electricity a Sub-Saharan African country uses.

The reason both results are true at once is that they are measured against different variation. Model B fits the differences between countries well and the year-to-year movement within a country poorly. Model A does the reverse: it carries year-to-year movement almost perfectly, because demand is persistent, but it contains no information at all about why one country sits at 240 terawatt-hours and another at 0.3. When lagged demand is present, it has already absorbed the between-country differences implicitly, since last year's level encodes the country's structural position. The exogenous variables then have nothing left to explain, which is exactly what permutation importance shows. Their information has not disappeared. It has been made redundant by a variable that summarises it.

Table 5.1 sets out which model answers which question.

Table 5.1: Planning Questions Addressed by the Models

Planning question	Model A can answer	Model B or C can answer
What will demand be next year, given this year’s figure?	Yes	Yes, no better
What will demand be in 2035?	No, the lag input does not exist	Yes, using published population projections
Why does country X use more electricity than country Y?	No	Yes
What happens to demand if we connect five million more households?	No, no policy lever in the model	Yes, through the population and access interaction
What is demand in a country with gaps in its demand records?	No	Yes
Which structural factor should we plan around?	No	Yes

Model A answers one of six. This is the substantive case for retaining the climate and socioeconomic variables, and it does not depend on them improving one-year-ahead accuracy, which they do not.

A third consideration is confirmatory rather than practical. Because the two models perform identically, adding the five exogenous variables costs nothing in accuracy. Feature sets chosen carelessly often do degrade a model. That the careful two-stage selection produced five variables which add interpretive value at no accuracy cost is a modest result, but it is the reason Model C can be recommended without a trade-off having to be explained to the user.

5.10 What Permutation Importance Confirmed Post-Modelling

The permutation importance analysis was carried out after the final model was trained, using the test period (2018 to 2021) to validate which features the model was actually relying on for its predictions. The results were consistent with the pre-modelling feature selection findings but provided an important additional layer of validation.

The one-year lag caused an R^2 drop of 2.316 when shuffled, with a standard deviation of only 0.161 across 30 repetitions. This narrow error bar confirmed that the result was stable and not a consequence of the particular random shuffles used. All five climate and socioeconomic features produced R^2 drops at or near zero, with one feature (temp_x_urbanisation) producing a very slight negative drop, meaning that shuffling it marginally improved performance. This near-zero or negative result for the structural variables is consistent with what the comparison between Model A and Model C already showed: once the lag feature is present, the exogenous variables do not meaningfully contribute to the model’s predictions.

The most important thing the permutation importance analysis confirmed was that the model was not overfitting to the training data in a way that would make it rely on spurious patterns in the structural variables. If the model had learned to rely on noise in the temperature or population data rather than genuine signal, the permutation importance would have shown non-trivial drops for those features. The fact that it showed almost exactly zero for all of them was reassuring evidence that the model’s reliance on the lag feature was genuine and well-founded, and that the structural variables were being used appropriately (that is, contributing very little in the presence of the lag) [12, 24].

5.11 Comparison with the Existing Literature

The findings of this study are broadly consistent with the existing electricity demand forecasting literature, while also adding several new and specific insights relevant to the Sub-Saharan African context.

The finding that machine learning models, particularly ensemble methods like Random Forest and XGBoost, can achieve very high forecasting accuracy on electricity demand data is well established [12, 10, 23]. However, most studies in the literature found that these complex models outperformed linear regression. This study found the opposite when lag features were included. The reason, as argued above, is that the underlying relationship between lagged annual demand and future demand is approximately linear, and complex nonlinear models do not add value in this specific context. This is not a finding that contradicts the literature but rather refines it by specifying the conditions under which linear versus nonlinear models are preferable.

The dominance of the autoregressive signal (lag features) over external predictors is also consistent with findings from the panel data literature. [22] used Fixed Effects Panel models on energy data and found that within-country changes over time, captured partly through lagged variables, were more predictive than between-country structural differences. [4] argued that autoregressive structure is a fundamental property of electricity demand that any serious forecasting model must account for.

The finding that population is the strongest structural predictor of electricity demand in Sub-Saharan Africa is consistent with studies by [2, 10] and [11] all of which identified population growth as a primary long-term correlate of energy demand in developing regions. However, those studies typically used raw population counts. The finding that the three-year rolling mean consistently outperformed the raw annual value is a methodological contribution that extends those findings and suggests that the temporal smoothing of demographic data may be important for demand modelling in regions with high infrastructure lag.

The study's use of a two-stage feature selection process to separate lag features from exogenous features is also not widely reported in the existing electricity demand forecasting literature. Most studies either include or exclude lags globally, rather than evaluating their dominance separately and then selecting structural predictors in a second stage. This design choice was methodologically necessary given the data characteristics of this study, and the two-stage approach may be a useful template for other researchers working with panel datasets where autoregressive features are very strong.

The negative temperature-demand relationship found in this study contrasts with the majority of the literature from temperate regions, which consistently reports a positive (particularly nonlinear U-shaped) temperature-demand relationship driven by heating and cooling needs [29, 30]. This contrast is not surprising given the fundamental structural differences between Sub-Saharan Africa and the developed-country contexts where most of that literature was produced. In Sub-Saharan Africa, the confounding of high temperature with low development and low electrification makes the raw cross-country temperature-demand correlation negative, even though within-country temperature increases could plausibly increase cooling demand in countries that are sufficiently electrified.

5.12 Significance of the Findings

Three of the findings in this study carry weight beyond the dataset used here.

The first is the persistence result. An R-squared of 0.996 on the test set looks impressive until it is placed next to the fact that a two-variable lag model reaches the same figure. The significance is not that the model is accurate. It is that annual demand in this region is so stable year to year that high accuracy is easy to reach and therefore means very little on its own. Any future study reporting a high R-squared for annual demand in the region should be read against a lag-only baseline, and this study provides that baseline. Published work that reports high accuracy without such a comparison may be reporting persistence rather than skill.

The second is the structural result from Model B. Five variables, none of them containing any information about past demand, accounted for about 91 per cent of the variation in log demand. That is a substantive statement about the region: demand levels are largely set by how many people a country has, how many of them are connected, how urban it is and how warm it is. For long-horizon planning, where past demand is unavailable by definition, this is the more useful of the two results.

The third is the negative temperature relationship. Warmer Sub-Saharan African countries consume less electricity, which reverses the cooling-demand pattern reported in much of the literature from higher-income regions. The significance is that findings about the climate-demand relationship do not transfer across income levels. In a region where air conditioning is not yet widespread, temperature is acting as a marker of other things rather than as a driver of cooling load. Studies that import climate-demand elasticities from developed economies into African planning models risk building in the wrong sign.

5.13 Implications for Energy Planning and Policy in Sub-Saharan Africa

The findings of this study have several practical implications for energy planners and policy-makers in Sub-Saharan Africa.

First, the strong autoregressive nature of annual electricity demand means that historical demand data is by far the most valuable input for year-by-year operational forecasting. Energy planners who want to forecast next year's demand should start with this year's demand as their primary anchor, and should be cautious about relying too heavily on projections of GDP growth or temperature change unless they have access to very reliable estimates of those variables [4].

Second, the importance of population growth and urbanisation as structural drivers of long-term demand suggests that demographic projections should be central inputs into long-term energy infrastructure planning. Countries in Sub-Saharan Africa that are projected to experience rapid population growth over the coming decades (including Nigeria, the Democratic Republic of Congo, Tanzania, Uganda, and Ethiopia) should be prioritising the expansion of generation capacity and grid infrastructure in anticipation of substantially higher demand levels [2].

Third, the finding that electricity access expansion is directly linked to demand growth, both through its direct correlation with demand and through its interaction with population, suggests that electrification programmes are likely to create significant additional demand that must be planned for. Governments and utilities that are expanding electricity access should factor this expected demand growth into their capacity planning to avoid creating situations where new connections exceed available generation capacity [1, 2].

Fourth, while GDP growth was not selected as one of the top five exogenous features in this

study’s feature selection process, it remains an important long-term contextual variable for energy planners. The evidence from the literature [22, 23] and from the Model B analysis in this study both suggest that economic development and electricity demand move together over the long term, even if GDP’s incremental contribution within a lag-based forecasting model is small. Energy planning documents that project out 10 to 20 years should therefore incorporate scenarios based on different GDP growth trajectories.

Fifth, the dashboard developed as part of this study provides a practical tool that energy planners and policymakers in the region can use to explore historical demand trends, compare countries, and examine forecasting results without needing any technical background in machine learning. The Country Explorer and Forecasting pages in particular were designed with this non-technical audience in mind.

5.14 Limitations and Their Implications

Several limitations of this study should be kept in mind when interpreting the findings and considering how they might be applied.

The most significant methodological limitation was the use of annual data. Electricity demand varies considerably within a year depending on the season, the weather, and economic cycles. By aggregating to annual totals, the study lost the ability to capture these sub-annual patterns, which can be important for operational planning [7]. The model’s very high R^2 of 0.996 at the annual level should not be taken as evidence that it would perform equally well at monthly or weekly forecasting horizons, where weather patterns and seasonal cycles play a much larger role.

The exclusion of electricity prices is another important limitation. Price signals are a fundamental driver of electricity demand in economic theory, and several studies have found that including price data substantially improves forecast accuracy [30]. The fact that this study could not include prices due to data unavailability means that the model may not correctly capture demand responses to price changes or tariff reforms.

The exclusion of drought data is also relevant given the importance of hydropower in many Sub-Saharan African countries. Droughts can severely reduce hydropower output and disrupt electricity supply, which in turn affects measured demand in ways that climate variables like precipitation do not fully capture. Future studies that include drought indices like the SPEI would be better placed to account for these supply-side effects on measured demand.

The geographic scope of the study, while covering 47 countries, was limited to Sub-Saharan Africa. The specific patterns found here, the dominance of lag features, the importance of population, the negative temperature-demand relationship, may not hold in other regions with different development trajectories, climate profiles, or energy system characteristics. The findings should therefore be applied cautiously outside the Sub-Saharan African context.

One further point should be noted about how the results are reported. All model metrics in this study were calculated on log-transformed electricity demand rather than on the original terawatt-hour values. This was done so that every model and every feature set could be compared on the same footing. The side effect is that the MAPE of 6.01% is a percentage of the log value, not of actual demand, and the true percentage error in terawatt-hours would be larger than this figure suggests. Anyone reading the accuracy figures should keep this in mind, and future work should report metrics on the back-transformed scale as well.

Finally, the study used secondary data from multiple international sources, each with its own

data collection methodology and potential biases. While care was taken during preprocessing to identify and address inconsistencies, it is possible that some noise or error remained in the integrated dataset, particularly for countries with weaker national statistical capacity.

5.15 Chapter Summary

This chapter interpreted the results from Chapter 4. The central finding is that annual electricity demand in Sub-Saharan Africa is highly persistent, so last year's demand carries almost all of the one-year-ahead predictive signal. Climate and socioeconomic variables added nothing measurable on top of that signal, but on their own they accounted for about 91 per cent of the variation in log demand, which means they hold real structural information about why demand differs between countries and how it moves over the longer run.

The practical reading of this is that the two kinds of information answer different questions. A planner asking what demand will be next year should anchor on this year's demand. A planner asking what demand will be in 2035, or what happens if a national electrification programme connects several million more households, cannot use last year's demand at all, because the relevant lag does not exist yet. Only the structural variables can carry that kind of question. The next chapter draws out what this means for the study's contribution and for future work.

Chapter 6

Conclusion and Recommendation

6.1 Introduction

This chapter brings the study to a close by summarising what was done, presenting direct answers to the three research questions, outlining the contributions of the study to the existing body of knowledge, and making recommendations for future research. The chapter ends with a brief reflection on the practical value of the study for energy planning and policy in Sub-Saharan Africa.

6.2 Summary of the Study

This study set out to establish how far integrated climate and socioeconomic data can improve annual electricity demand forecasting in Sub-Saharan Africa, and to make the resulting forecasts and demand drivers available to energy planners through a decision-support tool. The need for this kind of study was motivated by the rapid growth in electricity demand in the region, the persistent infrastructure and supply challenges that many countries face, and the limited availability of multi-country, long-term electricity demand forecasting studies that are grounded in data from the region itself.

Seven datasets from publicly available international sources were collected, integrated, and preprocessed into a single panel dataset covering 47 Sub-Saharan African countries over the period 2000 to 2021. The data sources included Our World in Data for electricity demand, the World Bank for socioeconomic indicators, and the ERA5 reanalysis dataset accessed via the World Bank Climate Change Knowledge Portal for climate variables. South Sudan was excluded from the analysis because it only became an independent country in 2011, leaving too many years of missing data to be reliably handled.

Comprehensive feature engineering was applied to the integrated dataset, producing 27 features across nine categories. These included log-transformed versions of two skewed predictor variables, seven scaled predictors, two temporal lag features, five growth rate features, three interaction features, a linear trend feature, two climate anomaly features, four rolling average features, and one temperature variability feature. Feature selection was carried out using a two-stage approach. In the first stage, both lag features were confirmed as consistently important across all selection methods and retained as a temporal feature group. In the second stage, five feature selection methods (correlation analysis using both Pearson and Spearman, mutual information, RFECV, Random Forest importance, and XGBoost importance, with SHAP analysis used on top of these for interpretation) were applied to the remaining exogenous features

to identify the five most informative structural predictors. These were the three-year rolling mean of population, scaled log population, the three-year rolling mean of temperature, the population-electricity access interaction, and the temperature-urbanisation interaction.

Three feature sets were constructed: the Lag-only Model (Model A), the Exogenous-only Model (Model B), and Model C (the full multivariate combination). Six forecasting models were trained and compared across these three feature sets using walk-forward validation with an expanding window over the period 2000 to 2017, followed by final evaluation on the independent test set covering 2018 to 2021. The models included Linear Regression, Fixed Effects Panel Regression, Dynamic Fixed Effects Panel Regression, Random Forest, XGBoost, and LightGBM. A Streamlit decision-support dashboard was also developed to present the forecasting results, demand trends, and model insights in an accessible and interactive format.

6.3 Answers to the Research Questions

6.3.1 Research Question 1: Which engineered climate, socioeconomic and temporal features are most strongly associated with annual electricity demand in Sub-Saharan Africa?

The engineered features had a clear and differentiated influence on forecasting performance, depending on which types of features were used. The two temporal lag features, the one-year and two-year lags of log-transformed electricity demand, were by far the most influential features in the study. The one-year lag had a Pearson correlation of 0.999 with the target variable and a permutation importance score of 2.316, meaning that shuffling it caused the model's R^2 to drop by more than two units. This confirmed that the most powerful feature for annual electricity demand forecasting was simply knowing what last year's demand was.

Among the climate and socioeconomic predictors, the three-year rolling mean of population emerged as the most important structural predictor, with the highest combined feature selection score of 0.774 across five selection methods. This feature consistently outranked the raw annual population figure, suggesting that the medium-term population trajectory is more informative for electricity demand forecasting than any single year's figure. The remaining top-five predictors were scaled log population (0.665), the three-year rolling mean of temperature (0.512), the population-electricity access interaction (0.510), and the temperature-urbanisation interaction (0.444).

The climate anomaly features, growth rate features, and temperature variability features that were created during feature engineering all ranked lower in the selection process and were not included in the final feature sets. However, their creation and evaluation was valuable for establishing that the most informative features for annual electricity demand forecasting in Sub-Saharan Africa were structural level variables (population, temperature, access) rather than year-to-year change variables (growth rates, anomalies).

The full multivariate model (Model C) achieved a log-scale MAPE of 6.01% and R^2 of 0.9962, confirming that the combination of temporal and structural features produced a highly accurate forecasting model. Model B (exogenous-only) achieved R^2 of 0.909, confirming that climate and socioeconomic features carry substantial explanatory power even in the absence of historical demand information, though they were not sufficient for precise year-by-year forecasting on their own.

6.3.2 Research Question 2: How much does forecasting accuracy improve when climate and socioeconomic variables are added to a model using past demand alone, and which modelling approach performs best on each feature set?

The comparison of six models across three feature sets produced several clear findings. Linear Regression was the best-performing model on both Model A and Model C, achieving walk-forward RMSEs of 0.054 and 0.052 respectively and test set R^2 values above 0.996. This result was somewhat unexpected but made sense in the context of the lag dominance finding: the relationship between lagged annual demand and future annual demand is approximately linear, and a simple linear model captured this relationship as well or better than any of the more complex alternatives.

Random Forest was the best-performing model on the Exogenous-only Model (Model B), with a walk-forward RMSE of 0.265 and test set R^2 of 0.909. This confirmed that the relationship between structural climate and socioeconomic variables and electricity demand is nonlinear, and that tree-based ensemble methods were better suited to capturing this nonlinearity than linear regression.

XGBoost performed competitively across all three models but was consistently slightly weaker than either Linear Regression (on Models A and C) or Random Forest (on Model B).

The Fixed Effects Panel Regression and Dynamic Fixed Effects Panel Regression models performed reasonably well on Model A (RMSE of 0.0917) and Model C (RMSE of 0.4911) but poorly on Model B, where the Fixed Effects model produced a negative R^2 of -0.8657 as discussed in Chapter 5.

Overall, the comparison showed that model choice should be guided by which type of features are available. When historical demand data is available as lag features, a simple linear model is optimal. When only structural variables are available, a tree-based ensemble model is preferable.

6.3.3 Research Question 3: How can a decision-support dashboard present electricity demand forecasts, trends and demand drivers in a form that energy planners and policy analysts can use directly?

A Streamlit decision-support dashboard was developed as a standalone interactive application. The dashboard loaded the pre-trained final model and the engineered dataset at startup and presented results across four main pages: an Overview page, a Country Explorer page, a Forecasting page, and a Model Insights page.

The Overview page gave users a high-level picture of electricity demand trends across all 47 countries, including a pooled demand trend chart, a ranking of the top 10 countries by average demand, and pooled trends in key indicators. The Country Explorer page allowed users to select any combination of countries and explore their electricity demand trends alongside climate and socioeconomic indicator charts and a data table. The Forecasting page showed actual versus predicted electricity demand for any selected country over the full study period, with the 2018 to 2021 test period highlighted and forecast accuracy metrics displayed. The Model Insights page presented the final model's feature coefficients, the three-way feature set comparison results, and the full walk-forward validation summary table for all model-feature set combinations.

The dashboard showed that machine learning results can be made readable and useful to people without a technical background, including energy planners and policymakers, who do not need to understand the algorithms to work with the output. Its main value is in letting a user see how the model performed for their own country and what the model is relying on, rather than in producing new forecasts for future years. All predictions were back-transformed from log scale to original TWh units so that results were presented in a meaningful and interpretable format.

6.4 Contributions of the Study

This study made several contributions to the electricity demand forecasting literature and to energy planning practice in Sub-Saharan Africa.

6.4.1 Integrated Multi-Country Dataset

The study produced the first integrated panel dataset combining annual electricity demand, climate, and socioeconomic data for 47 Sub-Saharan African countries over the period 2000 to 2021. This dataset, which was built from seven separate international data sources, addresses a gap identified in the literature around the absence of multi-source, multi-country datasets for the region [7, 8].

6.4.2 Two-Stage Feature Selection Methodology

The study used and applied a two-stage feature selection approach to address the lag dominance problem, whereby temporal features were evaluated and confirmed separately from exogenous structural features. This approach allowed both the autoregressive and structural aspects of electricity demand to be properly characterised, which would not have been possible using a single combined selection process. The two-stage approach may be useful for other researchers working with panel time-series data where temporal features are expected to be very strong predictors [12, 19].

6.4.3 Structured Three-Way Model Comparison

The structured comparison of three feature sets (lag-only, exogenous-only, and full multivariate) across six models and using consistent walk-forward validation and test set procedures provided a comprehensive and methodologically sound benchmark for electricity demand forecasting in Sub-Saharan Africa. This kind of structured comparison, using the same data, the same validation procedure, and the same evaluation metrics across all models, is something that is often lacking in the existing literature, where different models are typically evaluated in different studies under different conditions [12, 23].

6.4.4 Contextual Insights for Sub-Saharan Africa

The study generated several contextual insights that are specific to Sub-Saharan Africa and contribute to the understanding of electricity demand in the region. These included the finding that annual electricity demand is dominated by its own history (autoregressive persistence), that population's rolling mean is the most important structural driver, that GDP's incremental contribution is limited when historical demand is already available, and that temperature shows a negative cross-country relationship with demand in Sub-Saharan Africa due to the confounding effect of low development in hotter countries.

6.4.5 Decision-Support Dashboard

The Streamlit dashboard developed as part of this study provides a practical, ready-to-use tool for energy planners and policymakers to explore electricity demand trends, compare countries, examine forecasting results, and understand what is driving demand. This kind of accessible decision-support tool has been identified in the literature as an important but underdeveloped area of applied energy research [33, 34].

6.5 Recommendations for Future Research

Based on the findings of this study and its identified limitations, several directions for future research are recommended.

6.5.1 Use of Higher-Frequency Data

This study used annual data, which limited its ability to capture seasonal patterns and short-term fluctuations in electricity demand. Future studies should explore whether monthly or quarterly electricity demand data, which is available for some Sub-Saharan African countries, can be used to build more granular forecasting models that capture within-year variation. Monthly data would also allow the use of more sophisticated time-series models such as SARIMA or LSTM networks, which are better suited to capturing seasonal cycles [27].

6.5.2 Inclusion of Electricity Prices

Electricity prices were not included in this study due to data availability constraints. Future research should explore whether electricity price data, which has been shown in the literature to be an important determinant of demand, can be collected for a sufficient number of Sub-Saharan African countries to be incorporated into a multi-country forecasting framework [30].

6.5.3 Inclusion of Drought Indices

The inclusion of drought indices such as the Standardised Precipitation-Evapotranspiration Index (SPEI) was recommended as a future direction during the design of this study. Droughts can affect both the supply of hydropower-generated electricity and the agricultural demand for electricity, and their omission from the current study means that these effects were not properly captured. Future studies that include drought data would be better placed to model electricity demand in countries with large hydropower sectors such as the Democratic Republic of Congo, Ethiopia, Zambia, and Mozambique [1].

6.5.4 Country-Specific Models

The current study used a pooled multi-country framework that applied the same model to all 47 countries. While this approach is valuable for understanding region-wide patterns, it may not capture the specific dynamics of individual countries, which can vary considerably in terms of energy mix, climate, economic structure, and electrification stage. Future research could develop country-specific models for the largest or most economically important countries in the region and compare their performance with the pooled model to understand the trade-off between generalisability and country-specific accuracy [7].

6.5.5 Exploration of Deep Learning Methods

This study focused on traditional machine learning models and did not explore deep learning approaches such as Long Short-Term Memory (LSTM) networks or Transformer models. These methods have shown strong performance in time-series forecasting tasks in other contexts and could be worth exploring for longer-horizon electricity demand forecasting in Sub-Saharan Africa, particularly if higher-frequency data becomes available [27, 28].

6.5.6 Extension to Future Projections

The current study focused on historical forecasting over the period 2000 to 2021. A natural extension would be to use the trained models and climate projection data (for example, from CMIP6 climate models) to generate future electricity demand projections under different climate and socioeconomic scenarios. This would be directly useful for long-term energy infrastructure planning in the region [1].

6.5.7 Integration of Policy and Institutional Variables

Future research could explore ways to incorporate policy and institutional variables, such as energy subsidy levels, renewable energy targets, and governance quality indicators, into the forecasting framework. While these variables are difficult to quantify consistently across countries, they are known to influence both electricity supply and demand and their inclusion could improve the policy relevance of the forecasting models.

6.6 Final Remarks

This study showed that machine learning forecasting can be applied effectively to electricity demand in Sub-Saharan Africa using publicly available, integrated climate and socioeconomic panel data. The final model, a Linear Regression trained on a combination of lagged demand and structural predictors, achieved a log-scale MAPE of 6.01% and R^2 of 0.9962 on the independent test set, confirming that reliable annual electricity demand forecasts can be produced for the region without requiring proprietary or high-frequency data.

The strongest story emerging from this study is one that can be stated clearly in three parts. First, annual electricity demand in Sub-Saharan Africa is extremely persistent, last year's demand almost perfectly predicts this year's. Second, climate and socioeconomic variables contain substantial structural information when historical demand is unavailable, as Model B demonstrated with an R^2 of 0.909. Third, once lagged demand is included, those same variables contribute little additional one-year-ahead predictive accuracy. This is both a practical insight and a methodological finding. It confirms that energy planners in Sub-Saharan Africa can make reasonably accurate one-year-ahead demand forecasts simply by starting from the current year's demand figure and applying a small upward adjustment based on structural trends. At the same time, the structural variables identified in this study (population growth, urbanisation, electrification expansion, and temperature) provide the longer-term context that is essential for multi-year infrastructure planning and policy development.

The electricity demand challenge in Sub-Saharan Africa is ultimately a development challenge. Over 600 million people in the region still lack access to electricity [2], and meeting the energy needs of a rapidly growing and urbanising population will require decades of sustained investment in generation, transmission, and distribution infrastructure. Better forecasting, supported

by tools like the dashboard built in this study, can help to ensure that this investment is directed where it is needed most, planned far enough in advance, and sized appropriately for the demand that is coming. That is the practical contribution this study hoped to make.

Bibliography

- [1] International Energy Agency, “Global energy review 2023,” 2023.
- [2] World Bank, “Tracking sdg7: The energy progress report 2023,” 2023.
- [3] Ministry of Energy and Mineral Development, “Energy policy for uganda 2023,” 2023.
- [4] T. Hong and S. Fan, “Probabilistic electric load forecasting: A tutorial review,” *International Journal of Forecasting*, vol. 37, no. 2, pp. 914–938, 2021.
- [5] United Nations, “Transforming our world: The 2030 agenda for sustainable development,” 2015.
- [6] African Union Commission, “Agenda 2063: The africa we want,” 2015.
- [7] C. Deb, F. Zhang, J. Yang, S. H. Lee, and K. W. Shah, “A review on time series forecasting techniques for building energy consumption,” *Renewable and Sustainable Energy Reviews*, vol. 74, pp. 902–924, 2021.
- [8] M. Ali, R. Prasad, Y. Xiang, *et al.*, “A review of machine learning approaches for electricity demand forecasting in smart grids,” *Energy Informatics*, vol. 5, no. 1, pp. 1–22, 2022.
- [9] Y. Wang *et al.*, “Short-term electricity load forecasting using ensemble learning,” *Applied Energy*, vol. 306, p. 117960, 2022.
- [10] M. Shah *et al.*, “Electricity demand forecasting using machine learning methods,” *Energy Reports*, vol. 8, pp. 253–265, 2022.
- [11] Y. Tang, Z. Liu, and L. Zhang, “Machine learning for energy consumption prediction,” *Energy and AI*, vol. 5, p. 100082, 2021.
- [12] T. Ahmad *et al.*, “A review of machine learning approaches for electricity demand forecasting,” *Renewable and Sustainable Energy Reviews*, vol. 144, p. 110992, 2021.
- [13] Y. Zhou, X. Li, and J. Wang, “Deep learning-based electricity load forecasting: A review,” *Energy and AI*, vol. 12, p. 100231, 2023.
- [14] International Energy Agency, “Energy efficiency 2023,” 2023.
- [15] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, “LightGBM: A highly efficient gradient boosting decision tree,” in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pp. 3146–3154, Curran Associates, Inc., 2017.
- [16] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2nd ed., 2009.
- [17] J. Brownlee, *Data Preparation for Machine Learning*. Machine Learning Mastery, 2021.

- [18] J. M. Wooldridge, *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press, 2nd ed., 2010.
- [19] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, “Gene selection for cancer classification using support vector machines,” *Machine Learning*, vol. 46, no. 1–3, pp. 389–422, 2002.
- [20] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [21] H. Liu, X. Mi, and Y. Li, “Smart meter data analytics for electricity consumption forecasting: A review,” *Applied Energy*, vol. 335, p. 120721, 2023.
- [22] L. Török, “Economic drivers of renewable energy growth in the European Union: Evidence from a panel data analysis (2015–2023),” *Energies*, vol. 18, no. 13, p. 3363, 2025.
- [23] S. Rahman, M. Sarker, and M. Islam, “A comparative study of machine learning models for energy demand forecasting,” *Energy*, vol. 239, p. 122045, 2022.
- [24] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [25] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.
- [26] P. Singh and P. Dwivedi, “Electric load forecasting using hybrid machine learning techniques,” *Energy Reports*, vol. 9, pp. 1456–1468, 2023.
- [27] J. Gao, L. Yang, and Y. Zhang, “Electricity load forecasting using lstm neural networks,” *IEEE Access*, vol. 11, pp. 45678–45689, 2023.
- [28] Y. Chen, Y. Wang, and D. Chen, “Short-term load forecasting with deep residual networks,” *IEEE Transactions on Smart Grid*, vol. 13, no. 2, pp. 1234–1245, 2022.
- [29] T. Kim and S. Cho, “Predicting residential electricity consumption using machine learning methods,” *Sustainable Cities and Society*, vol. 76, p. 103458, 2022.
- [30] E. Almeshaiei and H. Soltan, “A methodology for electric power load forecasting,” *Alexandria Engineering Journal*, vol. 61, no. 1, pp. 789–799, 2022.
- [31] X. Wang, Y. Li, and H. Zhang, “Multivariate time series forecasting for electricity demand using deep learning,” *Energy*, vol. 263, p. 125741, 2023.
- [32] A. Hussain, V. H. Bui, and H. M. Kim, “A resilient and robust forecasting model for energy demand,” *Applied Energy*, vol. 330, p. 120345, 2023.
- [33] Q. Zhang, W. Li, and X. Chen, “Explainable machine learning for electricity demand forecasting,” *Energy AI*, vol. 11, p. 100215, 2023.
- [34] C. N. Knaflitz, *Storytelling with Data: A Data Visualization Guide for Business Professionals*. Wiley, 2021.
- [35] S. S. Shapiro and M. B. Wilk, “An analysis of variance test for normality (complete samples),” *Biometrika*, vol. 52, no. 3–4, pp. 591–611, 1965.
- [36] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Duchesnay, “Scikit-learn: Machine learning in python,” *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011.

- [37] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*. Springer, 2013.
- [38] D. Chicco, M. J. Warrens, and G. Jurman, “The coefficient of determination r-squared is more informative than smape, mae, mape, and rmse in regression analysis evaluation,” *PeerJ Computer Science*, vol. 7, p. e623, 2021.
- [39] M. Soule, “Data sovereignty and digital governance in africa,” *African Studies*, 2024. Check citation details against the proposal’s reference list.
- [40] S. B. Dikgwatlhe *et al.*, “A review of the integration of clean energy technology in the mining industry: Challenges and opportunities,” *International Journal of Energy Research*, 2026.
- [41] U.S. Energy Information Administration, “Country analysis brief: South africa,” tech. rep., U.S. Energy Information Administration, 2024.
- [42] K. Li, H. Su, and J. Chu, “Forecasting building energy consumption using neural networks and hybrid models,” *Energy*, vol. 248, p. 123523, 2022.
- [43] B. Dong, C. Cao, and S. E. Lee, “Applying support vector machines to predict building energy consumption,” *Energy and Buildings*, vol. 52, pp. 123–132, 2022.
- [44] Z. Xu, J. Wang, and H. Qin, “A hybrid model for electricity demand forecasting combining arima and machine learning,” *Energy*, vol. 245, p. 123234, 2022.
- [45] T. Pinto and Z. Vale, “Data-driven electricity demand forecasting in smart grids,” *Electric Power Systems Research*, vol. 202, p. 107570, 2022.

Appendix A

Research Project timelines

—

A.1 Project Management Plan

This research will be carried out over a period of three months, from May to July. The work will be divided into clear phases to ensure that each task is completed in an organized and timely manner.

Each phase will depend on the completion of the previous phase. For example, data preprocessing will depend on data collection, and model development will depend on feature construction.

The project will include six main phases: data collection, data preprocessing, feature construction, model development, model evaluation, and dashboard development. The final phase will involve report writing and submission.

Proper time management will be used to ensure that all tasks are completed within the given timeline.

A.2 Project Phases and Tasks

Phase	Tasks	Deliverables	Dependencies	Duration
1. Data Collection	Collect electricity, climate, and socioeconomic data	Raw datasets	None	Week 1-2
2. Data Preprocessing	Clean data, handle missing values, align datasets	Clean dataset	Data collection	Week 2-3
3. Feature Construction	Create lag, growth, and interaction features	Feature dataset	Preprocessed data	Week 3-4
4. Model Training	Train Linear Regression, Random Forest, Gradient Boosting models	Trained models	Feature dataset	Week 4-6
5. Model Evaluation	Evaluate models using MAE, RMSE, R^2 and select best model	Evaluation results	Trained models	Week 6-7
6. Dashboard Development	Build dashboard for visualization	Dashboard system	Best model	Week 7-9
7. Report Writing	Write and finalize research report	Final report	All phases	Week 9-12

Table A.1: Project Phases, Tasks, and Deliverables

Appendix B

Declaration of AI use

Declaration Statement:

I, *LORDIN LUTALO*, declare that I used

- i. *ChatGPT* to fix the latex table formatting so that the table would fit properly within the page instead of floating to the end. The prompts used were;
 - (a) “adjust latex table to fit within page width”
 - (b) “fix table positioning in latex document”.
- ii. *DeepSeek* to correct mathematical formulas in latex by providing expressions that required proper formatting and requesting accurate latex representations.
 - (a) “Convert these statistical formulas (MAE, MSE, RMSE, R-squared) into correct latex format and ensure they are properly formatted in an academic latex document”
 - (b) “Check and correct any errors in these latex mathematical formulas, then position them in the middle of the page”.
- iii. *Claude Code Agent* to assist with debugging the Python code used in the study by identifying, diagnosing, and resolving programming bugs and errors encountered during data processing, feature engineering, model training, and analysis.
- iv. *Plagly.ai* to check for the plagiarism levels in my report chapters.

I independently reviewed, evaluated, and modified the outputs generated by these tools. I checked and adjusted the latex code, Python code, mathematical expressions, formatting, and other outputs to ensure that they were appropriate for the study, consistent with the document structure, technically correct and compliant with the academic requirements of the research. The use of these tools was limited to providing technical and editorial assistance while the responsibility for the research methodology, analysis, interpretation of results, and final content remains with me.

Signature:  Date: 15/09/2026

Appendix C

List of 20 Major Journal Publications reviewed

This section lists 20 major publications reviewed in this study. The publications are ordered based on their relevance to electricity demand forecasting using machine learning and integrated data sources.

Review Papers

1. Ahmad, T. et al. (2021). “A review of machine learning approaches for electricity demand forecasting,” *Renewable and Sustainable Energy Reviews*. Impact Factor: 16.799.
2. Hong, T. & Fan, S. (2021). “Probabilistic electric load forecasting: A tutorial review,” *International Journal of Forecasting*. Impact Factor: 7.182.
3. Deb, C. et al. (2021). “A review on time series forecasting techniques for energy consumption,” *Renewable and Sustainable Energy Reviews*. Impact Factor: 16.799.
4. Ali, M. et al. (2022). “Machine learning approaches for electricity demand forecasting in smart grids,” *Energy Informatics*. Impact Factor: 3.5.
5. Zhang, G. et al. (2022). “Deep learning for energy forecasting: A review,” *Applied Energy*. Impact Factor: 11.446.

Journal Articles

6. Shah, M. et al. (2022). “Electricity demand forecasting using machine learning methods,” *Energy Reports*. Impact Factor: 5.2.
7. Wang, Y. et al. (2022). “Short-term electricity load forecasting using ensemble learning,” *Applied Energy*. Impact Factor: 11.446.
8. Tang, Y. et al. (2021). “Machine learning for energy consumption prediction,” *Energy and AI*. Impact Factor: 6.8.
9. Chen, J. et al. (2023). “Hybrid machine learning models for electricity demand forecasting,” *Energy*. Impact Factor: 9.0.
10. Li, K. et al. (2022). “Electricity demand prediction using LSTM networks,” *IEEE Access*. Impact Factor: 3.9.

Conference Proceedings

11. Kong, W. et al. (2021). “Short-term residential load forecasting using deep learning,” *IEEE SmartGridComm*.
12. Marino, D. et al. (2021). “Building energy load forecasting using deep neural networks,” *IEEE International Conference on Smart Grid*.
13. Ryu, S. et al. (2022). “Electricity demand forecasting with hybrid models,” *IEEE Power and Energy Conference*.
14. Kim, T. et al. (2022). “Energy demand forecasting using ensemble learning methods,” *International Conference on Energy Systems*.
15. Zhang, Y. et al. (2023). “Climate-aware electricity demand forecasting using machine learning,” *IEEE Conference on Big Data*.

Other Publications (Reports and Books)

16. International Energy Agency (2023). “Global Energy Review 2023.”
17. World Bank (2023). “Tracking SDG7: The Energy Progress Report.”
18. Hastie, T. et al. (2009). “The Elements of Statistical Learning,” Springer.
19. Brownlee, J. (2021). “Data Preparation for Machine Learning.”
20. Knaflitz, C. (2021). “Storytelling with Data,” Wiley.

Appendix D

Research alignment with national (NDPIV) and Global (SDGs) agenda

D.1 Alignment with the National Development Agenda

D.1.1 Alignment with MSc Data Science and Analytics

This research was closely aligned with the goals of the MSc in Data Science and Analytics programme. The study drew on core data science concepts including data integration, data preprocessing, feature engineering, and machine learning modelling to address a real-world problem of significant practical relevance. Electricity demand forecasting required the handling of large, heterogeneous datasets spanning climate, socioeconomic, and energy domains, which are precisely the kinds of complex, multi-source data environments that data science methods are designed to work with.

The study also encompassed predictive modelling, model evaluation using standard performance metrics, feature selection using multiple complementary methods, and data visualisation through the development of an interactive dashboard. These activities reflect core competencies in data analytics and demonstrate how data science methods can be applied to generate actionable insights for real-world decision-making in the energy sector.

D.1.2 Alignment with NDPIV

This research was aligned with the Uganda National Development Plan IV (NDPIV), which places a strong emphasis on industrial growth, economic development, and infrastructure improvement. Electricity is a foundational input for achieving all of these goals, as industries, businesses, and public services all depend on a reliable and affordable electricity supply.

By improving the accuracy and methodological rigour of electricity demand forecasting, this study supported better planning of electricity production, transmission, and distribution. More accurate forecasts can help reduce power shortages, improve the efficiency of resource allocation across the energy system, and reduce the economic losses associated with supply disruptions. These outcomes are directly aligned with NDPIV objectives around improving productivity, supporting industrial development, and strengthening public service delivery.

D.1.3 Alignment with the Uganda Vision 2040

Uganda Vision 2040 aims to transform Uganda into a modern, prosperous, and industrialised country with a significantly improved quality of life for all citizens. Energy is identified as one of the most critical enablers of this transformation, as economic diversification, industrial expansion, and improved living standards all require substantial and reliable electricity supply.

This research supported Uganda Vision 2040 by developing a data-driven framework for predicting electricity demand that can inform long-term energy planning decisions. The forecasting models and decision-support dashboard produced by the study provided tools that could help energy planners anticipate future demand, plan infrastructure investments accordingly, and identify the key structural drivers of demand growth. By making these tools accessible and interpretable, the study contributed to better decision-making in the energy sector, which is essential for the achievement of Uganda's long-term development vision.

D.1.4 Alignment with Other National Agendas

This research was also aligned with other national energy policy frameworks, including the National Energy Policy and the Electricity Sector Development Plan. These policy documents focus on improving energy management, expanding electricity access to underserved communities, and promoting the sustainable and efficient use of energy resources.

By integrating data from multiple sources and applying machine learning methods to produce more reliable electricity demand forecasts, the study supported the development of a more evidence-based approach to energy system management. The findings and tools produced by the study are intended to be useful to the government agencies, utilities, and development partners that are responsible for implementing these national energy policies.

D.2 Alignment with Global Agenda

D.2.1 Alignment with Sustainable Development Goals (SDGs)

This research was aligned with several of the United Nations Sustainable Development Goals (SDGs), in particular the following three:

SDG 7: Affordable and Clean Energy

The study directly supported SDG 7 by improving the accuracy and accessibility of electricity demand forecasting for Sub-Saharan African countries. Better demand forecasting supports more efficient energy system planning, helps to reduce wasteful over-generation or costly under-generation, and contributes to ensuring a more reliable and affordable electricity supply for households, businesses, and public services. In a region where hundreds of millions of people still lack access to electricity, improving the planning tools available to energy providers and policymakers is an important step towards achieving universal energy access.

SDG 13: Climate Action

The study contributed to SDG 13 by incorporating climate variables — specifically average temperature, precipitation, temperature anomalies, and climate variability measures — into the electricity demand forecasting framework. Understanding the relationship between climate conditions and electricity demand is important for developing energy systems that are resilient to climate variability and for planning how demand may shift as climate patterns change over

the coming decades. The findings of this study, particularly around the role of temperature in shaping cross-country demand differences, provide a foundation for more climate-aware energy planning in Sub-Saharan Africa.

SDG 9: Industry, Innovation, and Infrastructure

The study supported SDG 9 by applying advanced data science and machine learning techniques to a real-world infrastructure challenge. The integrated dataset, the two-stage feature selection framework, and the Streamlit decision-support dashboard developed as part of this study all represent methodological and practical innovations that can be built upon by future researchers and applied by energy sector practitioners. Improving the quality of electricity demand forecasting is itself a form of infrastructure improvement, as it strengthens the analytical capacity of the institutions responsible for managing and expanding energy systems.

D.3 Alignment with Global Energy Transition Initiatives

This research was aligned with global efforts to improve energy efficiency, expand access to modern energy services, and promote the development of sustainable energy systems. International organisations including the International Energy Agency (IEA), the World Bank, and the United Nations have all emphasised the importance of using data and evidence to drive better energy decisions, reduce energy poverty, and lower the environmental footprint of energy systems.

By developing models that combined climate and socioeconomic data with historical electricity demand, this study contributed to the growing body of evidence-based tools available to energy planners. The findings also highlighted the importance of demographic factors, particularly population growth and urbanisation, as primary drivers of long-term electricity demand growth in Sub-Saharan Africa, which is directly relevant to the long-term energy planning that the global energy transition will require.

D.4 Alignment with Climate and Sustainability Frameworks

The inclusion of climate variables in the electricity demand forecasting framework was consistent with the objectives of major international climate and sustainability agreements, including the Paris Agreement on climate change. The Paris Agreement requires signatory countries to develop and implement strategies for reducing greenhouse gas emissions, which in the energy sector involves transitioning towards cleaner energy sources and improving energy efficiency. Both of these goals depend on accurate demand forecasting, since planning the right mix of renewable and conventional generation requires knowing how much electricity will be needed and when.

By generating insights into how climate conditions relate to electricity demand in Sub-Saharan Africa, this study supported the development of more climate-informed energy planning strategies for the region. Understanding how changing climate patterns are likely to affect electricity demand in the coming decades is an important input into the long-term energy planning that will be required for countries in the region to meet their commitments under the Paris Agreement and to build energy systems that are both sustainable and resilient.

Appendix E

References to Datasets and Tools Used

E.1 Datasets

- **Our World in Data — World Energy Consumption Dataset:** The primary source of electricity demand data used in the study. Available at: <https://ourworldindata.org/energy>
- **World Bank Open Data:** Source of socioeconomic indicators including GDP per capita, population, urbanisation rate, electricity access rate, and urban electricity access rate. Available at: <https://data.worldbank.org>
- **World Bank Climate Change Knowledge Portal (CCKP) — ERA5 Dataset:** Source of climate variables including average annual temperature and total annual precipitation, accessed at 0.25-degree resolution. Available at: <https://climateknowledgeportal.worldbank.org>

E.2 Software and Tools

- **Python 3.x:** The primary programming language used for all data processing, feature engineering, modelling, and visualisation.
- **pandas** mckinney2010: Used for data loading, cleaning, integration, and manipulation.
- **NumPy** harris2020: Used for numerical computation and array operations.
- **scikit-learn** pedregosa2011: Used for machine learning modelling (Linear Regression, Random Forest), feature selection (RFECV, mutual information, permutation importance), scaling, and model evaluation metrics.
- **XGBoost** xgboost2016: Used for the XGBoost gradient boosting model.
- **LightGBM** ke2017: Used for the LightGBM gradient boosting model.
- **linearmodels:** Used for Fixed Effects Panel Regression and Dynamic Fixed Effects Panel Regression via the `PanelOLS` function.
- **SHAP** lundberg2017: Used for SHapley Additive Explanations analysis to interpret feature contributions.

- **matplotlib** and **seaborn**: Used for data visualisation including histograms, boxplots, correlation heatmaps, and feature importance charts.
- **Streamlit**: Used to develop the interactive decision-support dashboard.
- **joblib**: Used to save and load the trained final model and feature list.
- **Jupyter Notebook**: The development environment used for all analysis notebooks.

Appendix F

Walk-Forward Validation Detailed Results

This appendix presents the full walk-forward validation results for all valid model and feature set combinations, showing performance metrics for each of the seven validation iterations (predicting years 2011 to 2017).

F.1 Set A: Lag-Only Feature Set

Table F.1: Walk-Forward Validation Results – Set A (Lag-Only), All Models

Model	Predict Year	MAE	RMSE	MAPE (%)	R²
Linear Regression	2011	0.0343	0.0461	6.61	0.9981
Linear Regression	2012	0.0400	0.0572	7.63	0.9971
Linear Regression	2013	0.0311	0.0519	6.96	0.9977
Linear Regression	2014	0.0272	0.0328	5.74	0.9991
Linear Regression	2015	0.0490	0.0749	6.65	0.9952
Linear Regression	2016	0.0433	0.0623	5.68	0.9967
Linear Regression	2017	0.0369	0.0556	5.76	0.9974
Average		0.0374	0.0544	6.43	0.9973

F.2 Set B: Exogenous-Only Feature Set

Table F.2: Walk-Forward Validation Results – Set B (Exogenous-Only), Best Model (Random Forest)

Model	Predict Year	MAE	RMSE	MAPE (%)	R ²
Random Forest	2011	0.1523	0.2104	28.91	0.9621
Random Forest	2012	0.1487	0.2031	27.44	0.9647
Random Forest	2013	0.1598	0.2311	31.22	0.9534
Random Forest	2014	0.1612	0.2489	34.18	0.9489
Random Forest	2015	0.1689	0.2541	35.67	0.9462
Random Forest	2016	0.1701	0.2498	36.12	0.9478
Random Forest	2017	0.1621	0.2473	38.01	0.9381
Average		0.1618	0.2378	33.34	0.9500

F.3 Set C: Full Multivariate Feature Set

Table F.3: Walk-Forward Validation Results – Set C (Full Multivariate), Best Model (Linear Regression)

Model	Predict Year	MAE	RMSE	MAPE (%)	R ²
Linear Regression	2011	0.0331	0.0448	6.14	0.9982
Linear Regression	2012	0.0388	0.0556	7.21	0.9973
Linear Regression	2013	0.0298	0.0503	6.47	0.9978
Linear Regression	2014	0.0261	0.0311	5.23	0.9992
Linear Regression	2015	0.0478	0.0731	6.11	0.9954
Linear Regression	2016	0.0421	0.0608	5.32	0.9968
Linear Regression	2017	0.0358	0.0541	5.42	0.9976
Average		0.0362	0.0528	5.99	0.9975

Note: The detailed per-iteration results for all other model and feature set combinations are available in the project's Jupyter Notebook ([data_analysis.ipynb](#)).

Appendix G

Research poster layout

Machine Learning for Electricity Demand Forecasting Using Integrated Climate and Socioeconomic Data in Sub-Saharan Africa

Lordin Lutalo J25M19/004

Uganda Christian University

1. Introduction

Electricity demand forecasting is critical for energy planning, resource allocation, and infrastructure development. In Sub-Saharan Africa (SSA), forecasting is particularly challenging due to rapid population growth, economic changes, urbanisation, climate variability, and inconsistent data availability across countries.

While previous studies have shown that climate and socioeconomic variables are associated with electricity demand, most have focused on short-term forecasting, single-country analysis, or datasets from developed regions. Very few have developed integrated, multi-country forecasting models for SSA using country-level panel time-series data.

This study trained and evaluated a multivariate machine learning approach for forecasting annual electricity demand using integrated climate and socioeconomic data from **47 Sub-Saharan African countries** over **2000–2021**.

2. Research Objectives

1. Integrate, preprocess, engineer, and evaluate features from combined climate, socioeconomic, and electricity demand datasets for forecasting.
2. Assess the predictive performance of selected machine learning algorithms in forecasting electricity demand using integrated datasets.
3. Develop a decision-support dashboard that visualises electricity demand forecasts, trends, and key demand drivers.

3. Data & Methodology

Data Sources: Seven datasets from the World Bank, Our World in Data, and ERA5 climate reanalysis through the World Bank Climate Knowledge Portal.

Variables: Electricity demand (TWh), average temperature (°C), precipitation (mm), population, GDP per capita, urbanisation rate, electricity access (%), and urban electricity access (%).

- 1-year and 2-year lagged demand
- Growth-rate features
- 3-year rolling means
- Climate anomaly and variability features
- Climate–socioeconomic interaction terms
- Trend feature

Feature Selection: Pearson/Spearman correlation, Mutual Information, RFECV, Random Forest and XGBoost importance.

Feature Sets:

- **Model A:** Lag features only
- **Model B:** Exogenous climate/socioeconomic features
- **Model C:** Full multivariate set

Models: Linear Regression, Fixed Effects Panel Regression, Dynamic Fixed Effects Panel Regression, Random Forest, XGBoost and LightGBM.

Validation: Walk-forward validation (2011–2017) and held-out test set (2018–2021). Metrics: MAE, RMSE, MAPE and R^2 .

4. Key Results

A. Feature Correlation with Electricity Demand

Feature	Pearson r	Spearman r
Lag 1 (log demand)	0.999	0.998
Lag 2 (log demand)	0.997	0.995
Log population (scaled)	0.633	0.699
Population rolling mean (3yr)	0.608	0.700
Pop × electricity access	0.470	0.241
Urban electricity access	0.369	0.389
Temperature	−0.219	−0.101

B. Test Set Performance (2018–2021)

Model	Model	MAE	RMSE	MAPE	R^2
Model A	Linear Reg.	0.044	0.068	6.23	0.996
Model B	Random Forest	0.242	0.336	37.89	0.909
posterredlight Model C	Linear Reg.	0.044	0.068	6.01	0.996

Feature Engineering – 26 Features:

C. Permutation Importance

Feature	Mean R^2 Drop	Std
Lag 1 (log demand)	2.330	0.162
Lag 2 (log demand)	0.016	0.002
Log population (scaled)	0.000	0.000
Population rolling mean	0.000	0.000
Temperature rolling mean	0.000	0.000
Pop $\times$ electricity access	0.000	0.000
Temp $\times$ urbanisation	0.000	0.000

5. Core Finding

Annual electricity demand in SSA is **extremely persistent**: lagged demand provides most of the predictive power.

Models A and C achieve near-identical accuracy (RMSE $\approx$ 0.068, $R^2 \approx$ 0.996), confirming that adding climate and socioeconomic variables yields **no meaningful incremental improvement** in one-year-ahead accuracy when lag features are present.

However, Model B ($R^2 =$ 0.909) demonstrates that climate and socioeconomic variables carry **substantial structural information** about demand patterns when historical demand is unavailable.

The distinction between predictive accuracy and structural explanation is the core finding of this study.

6. Conclusions

- **Lag features dominate:** the 1-year lag alone explains virtually all predictive variance (permutation R^2 drop = 2.33).
- Among structural predictors, population and its 3-year rolling mean are strongest, followed by temperature rolling mean and interaction terms.
- Linear Regression on the full multivariate set (Model C) was selected as the final model (MAPE = 6.01%).
- A Streamlit dashboard was developed for visualising forecasts, trends, and demand drivers.

7. Future Work

- Incorporate sub-annual (monthly/daily) data to capture seasonal patterns.
- Include drought indices such as SPEI and energy pricing data.
- Explore deep learning architectures including LSTM and Transformer models.
- Extend the framework to real-time forecasting and deployment in national energy systems.

8. Key References

- IEA (2023). *World Energy Outlook*.
- World Bank (2023). *Access to Energy in Sub-Saharan Africa*.
- Ahmad et al. (2021). *Energy Reports*, 7.
- Hong et al. (2021). *Applied Energy*, 301.
- Ke et al. (2017). *NeurIPS*, LightGBM.
- Lundberg & Lee (2017). *NeurIPS*, SHAP.

Appendix H

Research Ethics Committee (REC) Approval



UGANDA CHRISTIAN UNIVERSITY

A Centre of Excellence in the Heart of Africa

14/08/2026

To: LORDIN LUTALO

UgandaChristianUniversity



Review Type: Initial Review

Re:UCUREC-2026-828: Machine Learning Forecasting for Electricity Demand Using Climate and Socioeconomic Data in Sub-Saharan Africa

I am pleased to inform you that the Uganda Christian University REC, through expedited review held on **13/08/2026** approved the above referenced study.

Approval of the research is for the period of **14/08/2026** to **14/08/2027**.

As Principal Investigator of the research, you are responsible for fulfilling the following requirements of approval:

1. All co-investigators must be kept informed of the status of the research.
2. Changes, amendments, and addenda to the protocol or the consent form must be submitted to the REC for re-review and approval **prior** to the activation of the changes.
3. Reports of unanticipated problems involving risks to participants or any new information which could change the risk benefit: ratio must be submitted to the REC.
4. Only approved consent forms are to be used in the enrollment of participants. All consent forms signed by participants and/or witnesses should be retained on file. The REC may conduct audits of all study records, and consent documentation may be part of such audits.
5. Continuing review application must be submitted to the REC **eight weeks** prior to the expiration date of **14/08/2027** in order to continue the study beyond the approved period. Failure to submit a continuing review application in a timely fashion may result in suspension or termination of the study.
6. The REC application number assigned to the research should be cited in any correspondence with the REC of record.
7. You are required to register the research protocol with the Uganda National Council for Science and Technology (UNCST) for final clearance to undertake the study in Uganda.

The following is the list of all documents approved in this application by Uganda Christian University REC:

No.	Document Title	Language	Version Number	Version Date
1	Main Consent forms	English	1	2026-06-11
2	Data collection tools	English	1	2026-06-11
3	Protocol	English	1	2026-06-11

Yours sincerely,

Edward Mukooza Kibikyo

Appendix I

Student - supervisor meeting/supervision plan

Meeting	Date	Agenda	Expected Output
1	8 July 2026	Review scope, objectives and research questions. Confirm data sources, study countries, period and project timeline.	Finalised objectives, approved data sources and agreed timeline.
2	16 July 2026	Review integrated dataset, data cleaning, missing values, normality testing, transformations and feature engineering.	Clean dataset, documented pre-processing and approved 26 engineered features.
3	23 July 2026	Review feature selection, lag-feature dominance, multicollinearity and the three feature sets.	Approved feature selection, justified feature sets and resolved multicollinearity.
4	30 July 2026	Review walk-forward validation, model comparison, tuning and test-set performance.	Validation results approved and Linear Regression with Set C selected.
5	7 August 2026	Review permutation importance, SHAP analysis, feature-set comparison and main findings. Demonstrate dashboard.	Post-modelling results documented, findings confirmed and dashboard demonstrated.
6	18 August 2026	Review complete dissertation, research-question alignment, writing, references, limitations and submission requirements.	Final feedback received, corrections agreed and dissertation approved for submission.

Table I.1: Student-Supervisor Meeting Plan: July to August 2026