# APPLICATION OF MACHINE LEARNING FOR SOLAR IRRADIANCE FORECASTING IN ZAMBIA

by

**EMMANUEL CHILUBA MULENGA**

submitted in accordance with the requirements
for the degree of

**MASTER OF ENGINEERING**

in the subject

**ELECTRICAL ENGINEERING**

at the

**UNIVERSITY OF SOUTH AFRICA**

SUPERVISOR: Dr. Bessie B. Monchusi
CO-SUPERVISOR: Prof M. Sumbwanyambe

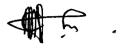
January 2026

# Declaration

I, Emmanuel C. Mulenga, declare that APPLICATION OF MACHINE LEARNING FOR SOLAR IRRADIANCE FORECASTING IN ZAMBIA is my own work and that all the sources that I have used or quoted have been indicated and acknowledged by means of complete references.

I further declare that I submitted the dissertation to the appropriate originality detection system, which is endorsed by Unisa and that it falls within the accepted requirements for originality.

I further declare that I have not previously submitted this work, or part of it, for examination at Unisa for another qualification or at any other higher education institution.



| | |
|---|---|
| | 14th January, 2026 |
| Signature | Date |

# Acknowledgement

I would like to begin by expressing my heartfelt gratitude to my wife, Josephine H. Mulenga, and mother, Brenda C. Mwila, whose unwavering love, encouragement and support have been the foundation of my academic journey. Their sacrifices, patience, and belief in me have sustained me through the most challenging phases of this work.

I am also profoundly grateful to my supervisors, Prof. Mbuyu Sumbwanyambe and Dr. Bessie B. Monchusi, for their exceptional guidance, mentorship, and insightful feedback throughout this research. Their expertise and consistent support have been instrumental in shaping the direction and quality of this thesis.

My sincere appreciation goes to my colleagues and mentors at the University of South Africa for fostering an engaging academic environment and for their continued encouragement.

Finally, I acknowledge the University of South Africa for providing the necessary facilities and resources that allowed the successful completion of this research.

Thank you for walking this journey with me.

**Emmanuel C. Mulenga**

# Abstract

Zambia has faced persistent energy shortages over the past decade due to heavy reliance on hydropower. Recent droughts, exacerbated by climate variability, have highlighted the vulnerability of this dependence and accelerated the country's diversification toward solar energy. Effective integration of solar generation into the national grid requires accurate forecasting of global horizontal irradiance (Global Horizontal Irradiance (GHI)) to mitigate intermittency and support efficient energy planning. Despite increasing investment in solar power, limited research exists on forecasting models tailored to Zambia's climatic conditions.

This study develops a data-driven GHI forecasting framework using ten years of local meteorological data, including GHI, temperature, humidity, precipitation, and wind speed, sourced from the Zambia Meteorological Department (ZMD) and Joint Research Center Photovoltaic Geographical Information System (JRC PVGIS). A systematic literature review guided by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework and bibliometric analysis examined global, regional, and local trends in solar irradiance forecasting. Based on the review, three widely applied machine learning models—Long Short-Term Memory (LSTM), Random Forest (RF), and Artificial Neural Networks (ANN)—were selected for evaluation. Feature selection employed Variance Inflation Factor (VIF) analysis and Least Absolute Shrinkage and Selection Operator (LASSO) regression, identifying temperature and humidity as the most relevant predictors.

The models were trained and tested using consistent data splits and evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and $R^2$. Results indicate that the ANN model achieved the highest predictive accuracy (MAE = 7.378 W/m$^2$, RMSE = 9.584 W/m$^2$, $R^2$ = 0.845), followed by RF (MAE = 9.845 W/m$^2$, RMSE = 12.374 W/m$^2$, $R^2$ = 0.715) and LSTM (MAE = 0.589 W/m$^2$, RMSE = 0.739 W/m$^2$, $R^2$ = 0.392). These findings demonstrate that reliable short-term GHI forecasting can be achieved using locally relevant predictors and accessible data.

The study provides practical value for energy planners, utilities, and policymakers seeking to enhance grid stability, optimize solar dispatch, and improve renewable energy integration in Zambia. It recommends adopting ANN-based forecasting tools and further exploring hybrid and region-specific approaches supported by enhanced local climate data collection.

***Keywords*— Machine Learning, Solar Irradiance, Forecasting, Renewable energy, Predictive Modeling**

# Table of Contents

# List of Figures

# List of Tables

# List of Abbreviations

**ML** Machine Learning

**RMSE** Root Mean Squared Error

**PV** Photovoltaic

**AI** Artificial Intelligence

**CNN** Convolutional Neural Network

**LSTM** Long Short-Term Memory

**ANN** Artificial Neural Networks

**RF** Random Forest

**MAE** Mean Absolute Error

**MSE** Mean Squared Error

**MAPE** Mean Absolute Percentage Error

**GHI** Global Horizontal Irradiance

**ZMD** Zambia Meteorological Department

**JRC PVGIS** Joint Research Center Photovoltaic Geographical Information System

**PRISMA** Preferred Reporting Items for Systematic Reviews and Meta-Analyses

**VIF** Variance Inflation Factor

**LASSO** Least Absolute Shrinkage and Selection Operator

**ERB** Energy Regulation Board

**PPA** Power Purchase Agreement

**HFO** Heavy Fuel Oil

**ISA** International Solar Alliance

**RL** Reinforcement Learning

**DNI** Direct Normal Irradiance

**DHI** Diffuse Horizontal Irradiance

**BHI** Beam Horizontal Irradiance

**MA** Moving Average

**ARIMA** Auto-Regressive Integrated Moving Average

**SES** Simple Exponential Smoothing

**AR** Autoregression

**sMAPE** Symmetric Mean Absolute Percentage Error

**TS** Tracking Signal

**MAD** Mean Absolute Deviation

**TS** Tracking Signal

**$R^2$** Coefficient of Determination

**MLR** Multiple Linear Regressionl

**CFE** Cumulative Forecast Error

**DSR** Design Science Research

**EDA** Exploratory Data Analysis

**ITCZ** Intertropical Convergence Zone

**RNN** Recurrent Neural Network

**GWR** Geographically Weighted Regression

**REFiT** Renewable Energy Feed-in Tariff

**REA** Rural Electrification Authority

**ANOVA** Analysis of Varaince

**MLP** Multilayer Perceptron

**NWP** Numerical Weather Prediction

**BiLSTM** bidirectional LSTM

**RPCA** Robust Principal Component Analysis

**rRMSE** Relative Root Mean Square Error

**SCA** Sine Cosine Algorithm

**FFNN** Feedforward neural network

**KNN** k-Nearest Neighbors

**DNN** Deep Neural Network

**NREL** National Renewable Energy Laboratory

**GBR** Gradient Boosting Regression

**RBF** Radial Basis Function

**SARAH** Surface Solar Radiation Data Set - Heliosat

**ERA5** ECMWF ReAnalysis 5th Generation

**BSRN** Baseline Surface Radiation Network

**SARIMA** Seasonal AutoRegressive Integrated Moving Average

# Chapter 1

# Introduction

This chapter provides an overview of the study, setting the foundation for the exploration of ML applications in the forecasting of solar irradiance within the Zambian renewable energy sector. It outlines the motivation for the research, presents the problem statement, and defines the research questions and objectives that guide the study. The chapter also highlights the significance of this research in addressing the challenges associated with renewable energy integration and effective solar power management in Zambia.

## 1.1 Background of the Study

Zambia's energy mix is predominantly composed of hydropower, which represents 83% of the total energy supply, reflecting a significant dependence on this renewable resource. Thermal generation contributes 7%, while diesel and Heavy Fuel Oil (HFO) each represent 2% and 3% of the mix, respectively, as shown in Figure 1.1. Additionally, other sources contribute 3%, and solar energy currently constitutes only 2% of the overall energy landscape [1]. This distribution underscores the need for diversification and increased investment in renewable resources, particularly solar, to improve energy security and sustainability in Zambia. Electricity access is also limited, with stark urban (85.7%) and rural (14.5%) disparities, affecting development [2].

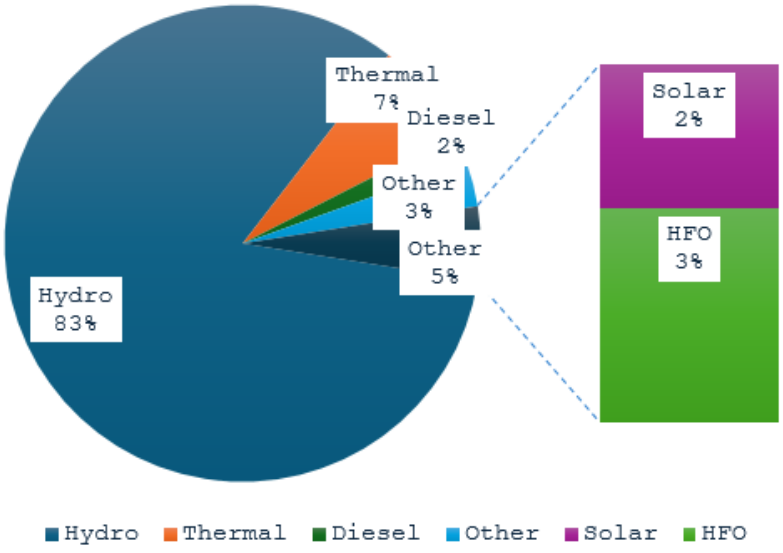



Figure 1.1: Power Generation Sources in Zambia

It should be noted that Zambia is endowed with significant solar energy resources, characterized by an average of 2,000 to 3,000 hours of sunshine per year and solar ir-

radiation of approximately 5.5 kWh/m$^2$ per day, making it one of the most favorable locations for solar energy generation in southern Africa [1]. Recognizing the critical need to diversify its energy portfolio, which has historically relied heavily on hydropower, the Zambian government has initiated various policies aimed at promoting the development of renewable energy, with a prominent focus on solar energy [3]. The Energy Regulation Board (ERB)'s approval of seven Power Purchase Agreement (PPA)s for solar and biomass projects, for example, demonstrates Zambia's commitment to moving toward affordable, sustainable, and reliable energy sources. Furthermore, the introduction of the biofuel blending pilot program marked a significant advancement in the integration of bioenergy into the energy mix, and in 2023, the Cabinet approved the ratification of the International Solar Alliance (ISA) France agreement, which supports financing and advocacy for the deployment of solar energy and is part of Zambia's commitment to the Paris Declaration [4].

Zambia's energy demand is predicted to increase by 121% by 2030 (5,422 MW) and 349% by 2050 (11,031 MW) due to an increasingly diversified economy, while solar power is predicted to reach 21% of total generation by 2030, as shown in Figure 1.2 [3]. This diversification is crucial, especially considering the country's vulnerability to climate variability, such as droughts, which severely affect water levels in reservoirs that support hydroelectric generation.

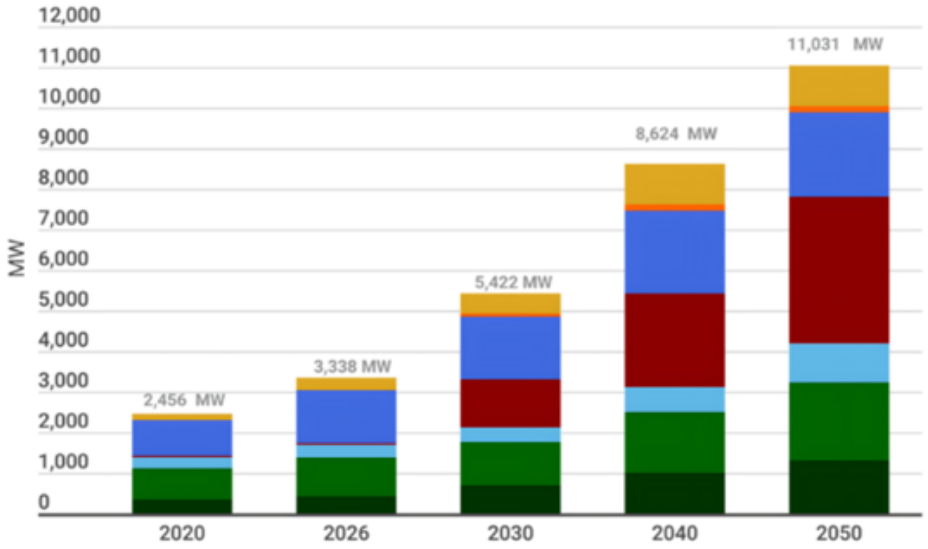



Figure 1.2: Zambia's Energy Demand forecast between 2020 and 2050

The ability to accurately forecast solar irradiance is vital for optimizing solar energy production and effectively integrating it into the national grid. Reliable solar irradiance forecasting helps energy providers in planning and managing solar power plants, allowing them to anticipate energy production with greater precision [5]. Improved forecasting capabilities enable grid operators to balance supply and demand effectively, thus minimiz-

ing disruptions and ensuring a stable energy supply, thereby transforming grid operations with accurate short-term energy forecasts.

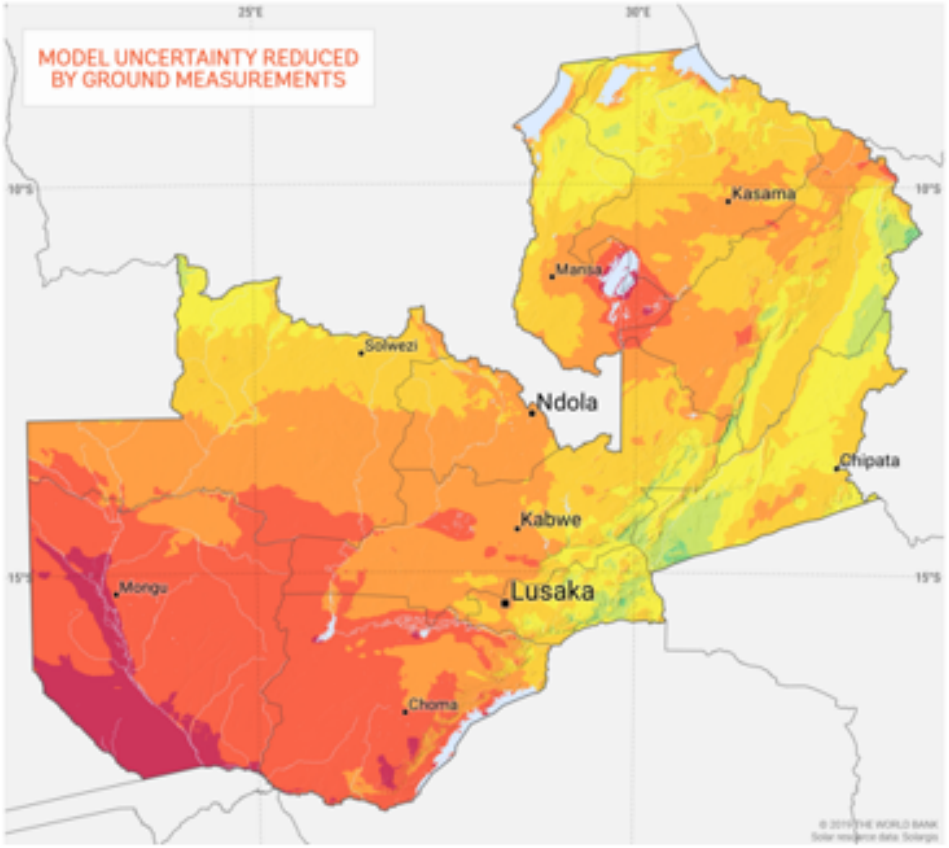



Figure 1.3: Zambia's Solar power potential

However, despite the abundant solar potential of Zambia as shown in Figure 1.3, the forecasting of solar irradiance faces significant challenges. Current methods rely predominantly on traditional statistical approaches, which often fail to capture the complex and dynamic nature of solar irradiance patterns influenced by local environmental factors [6]. These limitations can lead to inaccurate forecasts that ultimately hinder the effective deployment of solar energy resources.

In contrast, ML techniques present a powerful alternative to conventional forecasting methods. ML algorithms can analyze extensive datasets to identify patterns and relationships that remain obscured in traditional analyses [7]. By leveraging historical weather data, solar irradiance measurements, and various environmental factors, ML models can significantly enhance forecasting accuracy, providing a robust tool for energy management in Zambia [8]. In view of the aforementioned benefits of using ML models over traditional forecasting methods, this study aims to explore the application of machine learning in forecasting solar irradiance, specifically GHI, which represents the total solar radiation received per unit area on a horizontal surface. By evaluating the performance of selected ML models tailored to Zambia's climatic conditions, the study contributes

valuable insights into the role of data-driven approaches in renewable energy forecasting and supports Zambia's transition toward a more sustainable and diversified energy future.

## 1.2 Problem Statement

Despite Zambia's considerable potential for solar energy generation, the effective harnessing of this resource continues to face substantial challenges, largely due to the limitations of existing solar irradiance forecasting models [3]. Solar irradiance is the major and direct energy input into the terrestrial ecosystem and serves as the primary driving force behind physical, biological, and industrial systems [9]. The Sun's natural influence on climate and the earth's atmosphere makes solar forecasting critical for renewable energy planning. However, the variable absorption of sunlight by aerosols and clouds has hindered researchers' ability to accurately measure solar irradiance before it reaches the earth's surface [10]. This challenge is particularly pronounced in developing countries like Zambia, where solar irradiance measurements are scarce due to the high cost, maintenance needs, and calibration requirements of measuring equipment [11]. Current forecasting approaches predominantly rely on conventional statistical methods, which often fall short in capturing the inherently complex and highly dynamic nature of solar irradiance, particularly as influenced by Zambia's diverse and variable environmental conditions [6]. These traditional models typically assume linear relationships and are unable to fully account for non-linear atmospheric processes, localized weather patterns, and microclimatic variability that significantly affect solar energy availability (Gupta & Chauhan, 2019). Consequently, the resulting forecasts are frequently inaccurate, leading to suboptimal operational decisions by grid operators and energy planners [12]. These inaccuracies pose a critical barrier to effective integration of solar power into the national grid, ultimately affecting the country's broader efforts to diversify its energy mix, improve grid reliability, and achieve long-term renewable energy targets [13].

In response to these challenges, this research proposes the development of a ML-based solar irradiance forecasting model specifically tailored to Zambia's unique climatic and environmental conditions. By leveraging the advanced pattern recognition and adaptive learning capabilities of ML algorithms, the proposed model aims to provide more accurate and context-specific forecasts, thereby supporting more efficient energy management, improving the reliability of solar energy integration into the grid, and contributing to Zambia's transition towards a more resilient and sustainable energy future [3].

## 1.3 Research Questions

This study seeks to address the following key research questions:

1. What are the global and regional trends, advances, and research patterns in the application of ML techniques for solar irradiance forecasting, and which models and methodological approaches are most commonly used?

2. Which meteorological variables are most influential in accurately forecasting solar irradiance in Zambia, and how can statistical and ML-based feature selection techniques identify them?

3. How do selected ML models, identified from the literature review and adapted to Zambia's climatic conditions, perform in forecasting solar irradiance, and which model demonstrates the highest forecasting accuracy?

## 1.4 Objectives of the Study

General Objective: To evaluate existing ML models and select the most suitable for solar irradiance forecasting tailored to Zambia's environmental conditions, with the aim of improving forecasting accuracy and supporting effective energy management and planning.

Specific Objectives:

1. To examine global and local trends, advances, and research patterns in the application of ML techniques for solar irradiance forecasting, with a focus on identifying commonly used models and methodological approaches relevant to the Zambian context.

2. To identify the most influential meteorological variables for accurate solar irradiance forecasting in Zambia using statistical and ML-based feature selection techniques.

3. To evaluate and compare the performance of selected ML models, identified through the literature review and aligned with Zambia's climatic conditions, in forecasting solar irradiance, and to determine the most suitable model based on forecasting accuracy metrics.

## 1.5 Significance of the study

The findings of this study are anticipated to make significant contributions to both academic research and practical applications. By enhancing the accuracy of solar irradiance forecasting through the application of ML techniques, the study seeks to support the development of more efficient energy management strategies aimed at optimizing solar power generation. Improved forecasting accuracy has the potential to enhance grid stability by enabling energy providers to proactively manage supply and demand fluctuations [14]. Furthermore, the insights derived from this research are expected to provide valuable evidence to inform policymaking and guide investment decisions related to renewable energy infrastructure, thereby fostering the sustainable development of energy sector of Zambia. In addition, the broader implications of this study extend beyond the Zambian context, offering a potential methodological framework for other regions with similar climatic conditions and renewable energy challenges. By demonstrating the effectiveness of ML approaches in solar irradiance forecasting, the study may contribute to the wider

adoption of advanced forecasting techniques in diverse geographic and environmental settings, ultimately promoting more reliable and sustainable integration of solar energy into electricity grids globally.

## 1.6 Structure of the thesis

This thesis is structured into six key chapters, each contributing to the overall goal of exploring the application of ML in solar irradiance forecasting within the context of the Zambian renewable energy sector. Chapter 1 introduces the study by providing essential background information, highlighting the problem statement, and establishing the research questions and objectives. It emphasizes the importance of the study in addressing the challenges of renewable energy in Zambia. Chapter 2 delves into theoretical alternatives, exploring the foundational frameworks and key concepts related to ML, solar irradiance and forecasting techniques. It also provides a solid theoretical foundation to support the methodologies employed in the research. Chapter 3 presents a comprehensive review of the literature, analyzing existing studies on ML in solar irradiance forecasting, identifying research gaps and tracing the evolution of forecasting methods to establish the basis for the proposed contributions. Chapter 4 focuses on the research methodology, detailing the design, strategies, and techniques used in the study, including data collection methods, algorithm selection, and performance evaluation metrics to ensure transparency and reproducibility. Chapter 5 presents and discusses the research findings, interprets the performance of the evaluated ML models and highlights their practical implications for energy planning and management in Zambia. Finally, Chapter 6 provides a conclusion, summarizing key insights, reiterating the study's contributions, and suggesting future research directions, highlighting the critical role of ML advancing the sustainable energy transition of Zambia.

## 1.7 Scope and Limitations

This study will involve analyzing historical solar irradiance data, environmental factors, and weather conditions that are pertinent to the Zambian context. Although the primary objective is to provide a localized solution, it is important to acknowledge certain limitations inherent in this research. First, the availability and quality of comprehensive weather and irradiance data sets can pose challenges, as historical data are often limited or inconsistent. This may affect the robustness of the model and its ability to generalize the findings across different regions or climatic conditions. Second, the applicability of the developed forecasting model may be limited to areas with climatic characteristics similar to those found in Zambia. Although the model aims to improve the accuracy of the forecast within Zambia, applying the same techniques to other regions may require additional validation and adaptation. Lastly, the rapidly evolving nature of ML technologies means that new algorithms and techniques may emerge during or after the research

process, which could influence the relevance and performance of the proposed model over time. As such, ongoing research and iterative improvements will be necessary to maintain the effectiveness of the model in a dynamically changing energy landscape.

## 1.8 Chapter Summary

This chapter serves as an introduction to the study on the application of ML to solar irradiance forecasting in Zambia. It outlines the abundant solar resources of the country and the pressing need for effective forecasting methods to optimize solar power generation. The chapter discusses the vulnerabilities associated with Zambia's historical reliance on hydropower and the limitations of traditional forecasting approaches, which often fail to capture the dynamic nature of solar irradiance influenced by local environmental factors. To address these challenges, the chapter highlights the potential of ML as a more effective alternative to enhance forecast accuracy. The research questions focus on identifying suitable ML algorithms and evaluating their effectiveness in improving solar irradiance forecasting. The objectives aim to identify a customized an ML model and provide insights into energy management and policymaking. The next chapter will explore foundational theoretical concepts related to ML, solar irradiance and forecasting techniques. This chapter will establish the theoretical framework necessary to understand the methodologies and approaches used in this research, further illuminating their implications for Zambia's renewable energy forecasting.

# Chapter 2

# Theoretical Framework

This chapter will delve into the theoretical foundations of ML, solar irradiance and forecasting techniques. Understanding these foundational concepts is essential to appreciate the methodologies and approaches employed in this thesis. By highlighting these theoretical frameworks, the aim is to build a comprehensive understanding of how these elements interconnect and contribute to the primary goal of this study.

## 2.1 Machine Learning Overview

ML is a subset of Artificial Intelligence (AI) that enables systems to learn from data, identify patterns, and make decisions with minimal human intervention. ML can be defined as the ability of a computer algorithm to improve its performance on a specific task through experience [15]. Unlike traditional programming, where rules are explicitly coded, ML models learn from historical data, allowing them to make predictions or decisions based on new, unseen data. The field has gained significant traction in recent years, driven by the proliferation of data and advancements in computational power. This section will explore the fundamental concepts of ML. ML can be categorized into several types of algorithms, each serving different purposes, as shown in Figure 2.1

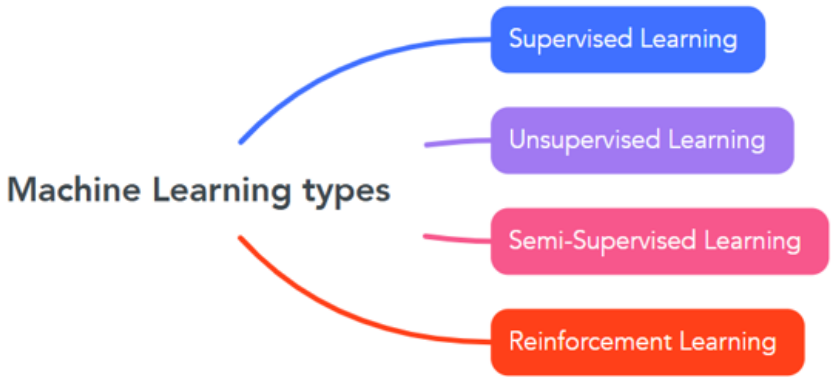



Figure 2.1: Classification of ML algorithms

### 2.1.1 Supervised Learning

Supervised learning is a type of ML where the model is trained on a dataset that includes both input features and the corresponding output labels [16]. The goal is to learn a mapping from inputs to outputs so that the model can predict the output for new, unseen data. To understand this, imagine you have a young child who loves playing with blocks.

You give them a mix of blocks of different shapes, like circles, triangles, and squares, and colors like blue, green, and red. The task is for the child to learn to separate these blocks based on either shape or color. If the sorting is by shape, the child will group all the round blocks together, all the triangular blocks together, and so on. But if the sorting needs to happen based on color, then all blue blocks go in one group, green in another, and so forth. Now, here is the question: how can the child know what "round" or "triangular" shapes are? Or how can they tell the difference between "blue" and "green"? Without prior knowledge, these terms mean nothing to them. This is where an adult comes in, guiding the child by pointing to each shape and color, helping them identify and label each feature.

In ML, the machine is like that child. To perform a sorting task, it needs a similar kind of guidance called training data. Training data is essentially past examples of different objects, labeled with information about their shape or color. These labels serve as the guide for the machine, just as an adult's explanations serve as guidance for the child. With enough labeled examples, the machine can start to recognize and categorize shapes and colors accurately. When training data is provided with labels like "round," "triangular," "blue," or "green," we call this supervised learning because the machine learns under supervision, like a child being taught by an adult. Supervised learning requires a substantial amount of labeled data, which can be time-consuming and expensive to obtain [16]. The quality and quantity of the training data significantly affect the performance of the model. Various algorithms are used in supervised learning, including linear regression for regression tasks, logistic regression, decision trees, support vector machines, and neural networks for classification task [17]. Common applications of supervised learning include classification tasks, which includes spam detection and image recognition; and regression tasks, such as predicting house prices and stock market trends [18].

### 2.1.2 Unsupervised Learning

Unsupervised learning involves training a model on data that does not have labeled outputs. The goal is to identify underlying structures or patterns within the data without any prior knowledge of the outcomes [19]. Unlike supervised learning, unsupervised learning does not require labeled data. This makes it particularly useful in situations where labeling is expensive or impractical. This type of learning is commonly known as a descriptive model, with the process often called pattern discovery or knowledge discovery [17]. A notable application of unsupervised learning is in segmenting customers based on shared characteristics. Evaluating the performance of unsupervised learning models can be more challenging than supervised learning since there are no ground truth labels [20]. Techniques such as silhouette scores or visual inspection of clusters are often used. Unsupervised learning can be used for feature extraction, where the model learns to represent the data in a way that captures its essential characteristics, which can then be used for

supervised learning tasks [20]. Some algorithms commonly used in unsupervised learning include K-Means Clustering, Hierarchical Clustering, and Gaussian Mixture Models. Although not commonly used directly in forecasting tasks, unsupervised techniques (e.g., clustering) can aid in data preprocessing by identifying structure in weather and irradiance data [21].

### 2.1.3 Semi-Supervised Learning

Semi-supervised learning involves training a model on a dataset that contains a small amount of labeled data and a large amount of unlabeled data. This approach leverages the strengths of both supervised and unsupervised learning to improve model performance [22]. The primary motivation for semi-supervised learning is the high cost and effort associated with labeling data. By using a small labeled dataset along with a larger unlabeled dataset, models can achieve better generalization and performance than they would with only labeled data [23]. Various techniques are employed in semi-supervised learning, including:

i. Self-training: The model is initially trained on the labeled data, and then it predicts labels for the unlabeled data. The most confident predictions are added to the training set iteratively.

ii. Co-training: Two or more models are trained in different views of the data, and they help each other by labeling the unlabeled data.

iii. Graph-based methods: These methods use the relationships between data points to propagate labels from labeled to unlabeled data.

Common applications include text classification, image classification, speech recognition and natural language processing tasks.

### 2.1.4 Reinforcement Learning

Reinforcement Learning (RL) is a prominent area of ML that focuses on training agents to make sequential decisions by interacting with their environment [20]. In this framework, an agent learns to optimize its actions based on feedback received in the form of rewards or penalties, thereby maximizing cumulative rewards over time. The learning process involves a critical balance between exploration; where the agent tries new actions to discover their effects; and exploitation, where it leverages existing knowledge to make informed decisions. Various algorithms, such as Q-learning and policy gradient methods, have been developed to facilitate this learning process, enabling applications across diverse fields including robotics, game playing, and autonomous systems [24]. Despite its potential, reinforcement learning poses challenges such as the need for extensive interaction with the environment and the complexity of designing effective reward structures

[25]. As a result, RL continues to be an active area of research, driving advancements in intelligent decision-making systems.

## 2.2 Insights into Solar Irradiance

Solar irradiance is the measure of solar power per unit area that reaches a given surface. It represents the intensity of sunlight at a particular moment, quantified in watts per square meter ($W/m^2$) [26]. It is often used when discussing instantaneous measurements, such as the intensity of sunlight hitting solar panels at a particular time. Solar irradiance varies based on factors such as the angle of the Sun, atmospheric conditions, and geographic location [27]. This measure is crucial for solar energy applications, as it directly affects the amount of energy that can be generated by solar panels or other solar collection devices [28]. By accurately assessing solar irradiance, one can optimize solar system designs, improve forecasting models for energy output, and better understand seasonal and hourly variations in sunlight [26]. It is worth noting that solar irradiance is a component of solar radiation. Solar irradiance consists of three components discussed in this section, through which it can be measured. These components, discussed in the following, are Direct Normal Irradiance (DNI), Diffuse Horizontal Irradiance (DHI), and GHI.

### 2.2.1 Direct Normal Irradiance

Also called beam irradiance, it is the part of the irradiance coming from the Sun that directly hits a surface without scattering [29]. It can also be represented as Beam Horizontal Irradiance (BHI) when measured on a horizontal plane. In other words, DNI represents the amount of radiation received per unit area on a surface perpendicular to the sun. The primary method of measuring DNI is with an instrument called a pyrheliometer [26]. Pyrheliometers typically employ thermopile sensors at the base of a light-collimating tube and glass window face, although they may also be constructed with another photosensitive element in place of the thermopile. If direct measurements of DNI are not available, it may be calculated via coplanar measurements of the diffuse and total radiation by devices with a 180° field of view, which is the incident angle between the collection plane and sun, which must also be known.

### 2.2.2 Diffuse Horizontal irradiance

DHI is the irradiance or radiation received by a horizontal surface that has been scattered or diffused by the atmosphere. It is the component of GHI which does not come from the Sun's beam [30]. DHI is typically measured with a pyranometer; however, in this case the direct light of the sun is blocked to remove the beam component of the radiation. The sun may be blocked by a ball or disc, which only removes the 5° cone around the sun and must utilize a tracker to continually shade only the sensor of the pyranometer [31]. The pyranometer may also be shaded by a horizontal or vertical shade band, the former requiring no tracker, the latter requiring a horizontal tracker; however, shade bands are

generally less accurate as they remove some of the diffuse light [30]. This diffuse light should be measured and corrected by a location- and time-dependent correction factor.

### 2.2.3 Global Horizontal Irradiance

GHI represents the total solar energy received on a flat, horizontal surface, such as the ground. This value accounts for both direct sunlight coming straight from the sun and diffuse sunlight scattered by the atmosphere [32]. GHI is a key parameter for assessing solar energy potential and is widely used in solar energy projects, weather prediction, and climate research. The primary device used to measure GHI is a pyranometer, which has a 180° hemispherical field of view. Pyranometers are specifically designed to capture solar radiation on a surface, with a critical feature being their cosine response to the angle of incoming light [31]. This means the sensor's output is directly related to the cosine of the angle at which sunlight strikes it. In other words, the pyranometer measures maximum radiation when the sun is directly overhead (at a 90-degree angle), and the recorded radiation decreases as the sun's angle shifts away from perpendicular, following a cosine relationship [29]. This cosine response ensures that the pyranometer delivers precise measurements across varying sun positions and atmospheric conditions, which is essential for reliable data in solar energy research, weather observation, and environmental monitoring. When direct GHI measurements are unavailable, it can be calculated using DNI and DHI using Equation 2.1.

$$\text{GHI} = \text{DHI} + \text{DNI} \cdot \cos(\theta_z) \tag{2.1}$$

Where:

- $\boldsymbol{\theta_z}$ is the solar zenith angle (angle between the vertical and the sun's position)

## 2.3 Forecasting Techniques

Forecasting is the process of making predictions about future events, conditions, or values based on historical data, current trends, and underlying patterns [33]. It is a crucial tool across various industries that allows predictions based on historical data and statistical or computational models. From finance to weather and manufacturing to healthcare, forecasting plays a vital role in decision-making, resource allocation, and risk management [33]. In general, forecasting techniques can be divided into statistical methods, ML methods, and hybrid approaches, each with unique strengths and applications. The primary objective of forecasting is to predict future values or trends based on past observations. Accurate forecasts enable organizations and stakeholders to make informed decisions, optimize resources, and mitigate risks [34]. For example, in energy systems, forecasting is essential to balance supply and demand, improve grid stability, and reduce operational costs. Key objectives of forecasting typically include:

- Predicting future demand or supply for resource planning
- Reducing uncertainty by providing probabilistic estimates
- Supporting strategic decision-making by anticipating market changes or environmental conditions
- Enhancing operational efficiency by optimizing resources based on predicted needs

The relevance of forecasting has grown as data has become more available and computational techniques have advanced, enabling increasingly complex models [34]. Forecasting techniques differ significantly based on the type of data used, the forecast timeframe, and the computational methods applied. They are generally grouped into qualitative and quantitative forecasting, discussed in the following sections.

### 2.3.1 Qualitative Forecasting

Qualitative forecasting is a subjective forecasting approach that relies on expert opinions, intuition, and experience instead of numerical data. It is particularly useful when historical data are limited or unreliable, or when forecasting for new products, emerging markets, or unpredictable situations [35]. Unlike quantitative methods that apply mathematical models to past data, qualitative forecasting uses expert judgment to predict future trends. It is commonly applied in early-stage research, strategic planning, and cases where data-driven forecasts are impractical [36]. While these methods lack statistical precision, they provide valuable insights, especially in areas where expert knowledge is crucial. Some qualitative forecasting methods are discussed below.

- *Delphi Method*: This is a structured process where a panel of experts provides forecasts independently. Their responses are collected, summarized, and shared with the group in several rounds. The process continues until a consensus, or a clear trend emerges. It helps reduce bias by keeping individual opinions anonymous [36]
- *Market Research*: This method involves gathering information directly from potential customers through surveys, interviews, or focus groups [35]. It helps us understand customer preferences, demand trends, and potential market shifts, which are then used to make forecasts
- *Expert Judgment*: In this approach, forecasts are based on the experience, intuition, and knowledge of industry experts [34]. It is useful in situations with limited historical data or in rapidly changing environments where expert insights provide valuable guidance

- *Scenario Analysis*: This involves developing multiple plausible future scenarios based on different assumptions (e.g., economic conditions, technology changes, policy shifts) and analyzing how these scenarios would affect outcomes [36]. It helps organizations prepare for a range of potential futures

### 2.3.2 Quantitative Forecasting

Quantitative forecasting uses numerical data and mathematical models to predict future trends. It relies on historical data and applies statistical and computational techniques to generate forecasts [34]. This method is suitable when there is sufficient historical data and clear patterns or relationships can be identified. There are basically two types of Quantitative forecasting, namely Time series forecasting and Casual or Explanatory forecasting [34].

**Time Series Forecasting**

Time series models are fundamental in quantitative forecasting, especially for data that follow identifiable patterns over time. By analyzing past observations, time series methods assume these patterns such as trends, seasonal cycles, and random fluctuations, will persist. Commonly used in time series include Moving Average (MA), exponential smoothing and Auto-Regressive Integrated Moving Average (ARIMA). These methods are briefly discussed below.

- *Moving Averages*: The MA method smooths time series data to reveal underlying trends by averaging a set of consecutive data points. This approach, often cited in foundational forecasting literature [37], is particularly effective in short-term forecasting for stationary data, where data values do not trend or show seasonality. MA can be computed using Equation 2.2.

$$MA = \frac{1}{n} \sum_{j=t-i}^{t} X_j \tag{2.2}$$

  Where:

  - $X_i$ = observed value at time $i$
  - $n$ = number of periods in the moving average

- *Simple Exponential Smoothing (SES)*: This improves upon the MA by assigning exponentially decreasing weights to past observations, prioritizing recent data points [33]. SES, calculated using Equation 2.3, is often used for nontrending, nonseasonal data, while Holt's and Holt-Winters [38] methods extend it to accommodate trends and seasonality.

$$F_{t+1} = \alpha X_t + (1 - \alpha) F_t \tag{2.3}$$

Where:

- $F_{t+1}$ is the forecast for the next period,
- $X_t$ is the observed value at time $t$,
- $\alpha$ is the smoothing constant, where $0 < \alpha < 1$; it determines the weight of recent observations.

Higher values of  increase the sensitivity of the model to recent changes, allowing faster adaptation to changes in trends or seasonal fluctuations, which is essential in forecasting highly variable data, such as solar irradiance [39].

- *Auto-Regressive Integrated Moving Average*: The ARIMA model is a flexible technique designed to handle non-stationary data by integrating three key components: Autoregression (AR), Integration, and MA. First introduced by Box and Jenkins [40], ARIMA models are extensively applied due to their effectiveness in capturing intricate time series patterns, making them a popular option for forecasting tasks like energy demand and solar irradiance forecasting [40]. ARIMA can be computed using Equation 2.4.

$$Y_t = \delta + \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t \tag{2.4}$$

Where:

- $Y_t$ is the observed value at time $t$,
- $\delta$ is the constant term,
- $\phi_i$ are the AR parameters,
- $\theta_j$ are the MA parameters,
- $\varepsilon_t$ is the error term (white noise).

### Causal (Explanatory) Forecasting

Causal models are grounded in the relationship between variables, forecasting a dependent variable based on changes in independent variables. In renewable energy forecasting, for example, causal models could assess how factors like temperature, humidity, or sunlight influence solar power output [41]. Regression analysis is a primary causal method, offering insights into variable relationships. Common methods in Casual forecasting include Linear regression, Multiple regression and Econometric models, briefly discussed below.

- *Linear Regression*: Simple linear regression analyzes the linear relationship between a dependent variable Y and a single independent variable X. This model,

foundational in predictive analytics [42], posits that changes in X directly impact Y, allowing forecasts based on trends in observable data. Linear regression is calculated using Equation 2.5.

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{2.5}$$

Where:

- $Y$ is the dependent variable (forecasted value),
- $X$ is the independent variable (predictor),
- $\beta_0$ is the intercept,
- $\beta_1$ is the slope coefficient, representing the strength of the relationship between $X$ and $Y$,
- $\varepsilon$ is the error term.

- *Multiple Linear Regressionl (MLR)*: When multiple factors influence the forecast variable, multiple linear regression offers a more comprehensive approach. This model considers several independent variables, making it well-suited for complex systems where various environmental or operational factors interact, such as in energy production modeling [42]. MLR is calculated using Equation 2.6

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon \tag{2.6}$$

Where:

- $X_1, X_2, \ldots, X_n$ are the independent variables (predictors),
- $\beta_1, \beta_2, \ldots, \beta_n$ are the corresponding regression coefficients,
- $\beta_0$ is the intercept,
- $\varepsilon$ is the error term.

- *Econometric models*: Econometric models are statistical tools used to forecast future trends by analyzing the relationships between economic variables. They combine economic theory with mathematical and statistical techniques to quantify how changes in factors like income, prices, or population affect outcomes such as energy demand or production [40]. These models are particularly useful for understanding complex systems where multiple variables interact, making them valuable for economic forecasting, policy analysis, and energy market predictions [43].

### 2.3.3 Evaluating Forecast Accuracy

Evaluating forecast accuracy is essential for assessing the reliability and performance of forecasting models. It helps identify the most effective models, fine-tune parameters, and optimize methods based on measurable metrics [33]. In time-series forecasting, accuracy evaluation ensures forecasts closely match observed values, boosting confidence in model reliability. Accurate forecasts are particularly critical in fields like energy, finance, and climate forecasting, where discrepancies can lead to significant economic, environmental, or operational impacts. For example, in renewable energy forecasting, precise solar irradiance forecasts are crucial for balancing power supply and demand. Various metrics are used to evaluate forecast accuracy, each providing valuable insights into model performance. Some common evaluation methods used to assess the accuracy of forecasting models include:

- *Mean Absolute Error*: MAE measures the average absolute differences between forecasted and actual values, providing a straightforward interpretation of forecast accuracy in the same unit as the observed data. MAE is particularly useful when the cost of errors is directly proportional to the error magnitude. MAE is calculated using Equation 2.7.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.7}$$

  Where:

  - $y_i$ is the observed value,
  - $\hat{y}_i$ is the predicted value,
  - $n$ is the total number of observations.

  MAE is easy to interpret and ideal for contexts where all forecast errors hold equal weight. However, it does not differentiate between over- and under-predictions, nor does it give more weight to larger errors.

- *Mean Squared Error (MSE)*: MSE, calculated using Equation 2.8, it calculates the average of squared differences between observed and forecasted values. By squaring the errors, MSE emphasizes larger errors, making it particularly suitable for applications where large deviations are more costly or significant.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2.8}$$

Since MSE amplifies the impact of large errors, it is ideal when high accuracy is prioritized for avoiding significant deviations. However, its value is in squared units, which may make interpretation less intuitive for some applications.

- *Root Mean Squared Error*: RMSE is the square root of MSE, providing an error measure that is interpretable in the same unit as the observed data. RMSE is often preferred when larger errors are particularly undesirable, as it emphasizes these deviations without the issue of squared units and can be calculated using Equation 2.9.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.9}$$

RMSE is particularly beneficial in applications where high error sensitivity is crucial, such as solar irradiance forecasting, where slight forecasting inaccuracies can disrupt energy production forecasts.

- *Mean Absolute Percentage Error (MAPE)*: MAPE expresses errors as a percentage of the actual values, making it highly interpretable and comparable across different scales. This metric is advantageous when evaluating forecast performance across datasets with different scales or units. MAPE is calculated using Equation 2.10.

$$\text{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{2.10}$$

MAPE is particularly valuable when model predictions need to be evaluated in relative terms, allowing stakeholders to understand forecast accuracy in a universally understandable format. However, MAPE can be sensitive to extremely small or zero actual values, potentially distorting results.

- *Symmetric Mean Absolute Percentage Error (sMAPE)*: sMAPE) is a variation of MAPE that addresses potential distortion in cases where observed values are near zero. By normalizing both actual and predicted values, sMAPE provides a balanced evaluation of over and under-predictions. sMAPE is calculated using Equation 2.11

$$\text{sMAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i|)/2} \tag{2.11}$$

This metric is especially useful for datasets with values close to zero or when it is essential to ensure that both over- and under-forecasting errors are equally penalized.

- *Coefficient of Determination ($R^2$)*: $R^2$ is a measure of the proportion of the variance in the observed data that is predictable from the independent variables. It provides an overall measure of model fit, with values ranging from 0 to 1, where values closer to 1 indicate a better fit. $R^2$ can be calculated using Equation 2.12.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{2.12}$$

  $R^2$ is particularly useful for understanding the proportion of variance explained by the forecast model and is widely used in regression-based forecasts.

- *Tracking Signal (TS)*: TS is used to detect bias in forecasting by comparing the cumulative forecast error to the Mean Absolute Deviation (MAD). A TS that falls within a range suggests an unbiased forecast, while values outside this range indicate a need for model adjustment. TS is calculated using Equation 2.13, with Cumulative Forecast Error (CFE) defined as the sum of all forecast errors over a period, and it is used to assess the bias in a forecasting model.

$$\text{TS} = \frac{\text{CFE}}{\text{MAD}} \tag{2.13}$$

  TS is a useful metric in time-series forecasting, where it is crucial to ensure that the model does not consistently over- or under-predict across time.

### 2.3.4 Forecasting Time Horizons

Forecasting time horizons represents the period for making predictions, influencing the choice of techniques, data, and model complexity. These horizons are categorized into short-term, medium-term, long-term, and ultra-long term, each serving specific operational or strategic purposes. Understanding these distinctions is essential in fields like energy, finance, and climate science, where accurate forecasts support daily operations and long-term planning, such as managing energy supply and demand or guiding infrastructure and policy decisions in renewable energy.

- *Short-term Forecasting*: Short-term forecasting typically covers periods ranging from minutes to a few days. It is predominantly used in operational decision-making where real-time or near-real-time data is available and rapid adjustments are necessary. In energy forecasting, short-term forecasts are crucial for managing daily load dispatch, grid stability, and real-time market trading [44]. Techniques commonly applied include time-series models such as ARIMA and ML models that can adapt quickly to new data patterns [45].

- *Medium-term Forecasting*: Medium-term forecasting spans weeks to a few months and is essential for tactical planning, particularly in industries such as manufacturing, finance, and energy, where decisions are made over several months to optimize processes. Medium-term forecasts help in planning for seasonal demand shifts, scheduling maintenance, and budgeting [16]. For example, in solar energy, medium-term forecasting can guide resource allocation during seasonal variations [46]. Techniques used here often involve hybrid models that combine statistical methods with ML for enhanced accuracy in the face of seasonal trends [33].

- *Long-term Forecasting*: Long-term forecasting encompasses periods from several months to years, aiding in strategic planning and investment decisions. Long-term forecasts are commonly used to inform infrastructure development, energy policy, and climate action planning. In renewable energy, these forecasts are essential for understanding capacity requirements and technology investments [47]. Long-term forecasting methods frequently incorporate econometric models, deep learning approaches, and scenario-based modeling to account for structural changes in systems over time [37].

- *Ultra-long-term Forecasting*: Ultra-long-term forecasting, spanning decades, is particularly relevant in fields where decisions have long-lasting impacts, such as urban planning, climate change adaptation, and energy transition [48]. Given the high uncertainty in such extended horizons, scenario analysis and probabilistic models are commonly employed to explore possible future states. For example, in climate science, ultra-long-term forecasts assess the potential impacts of various carbon reduction pathways [49]. These methods often require advanced simulation models that can incorporate complex interactions among variables over time [45].

## 2.4 Chapter Summary

This chapter presents a comprehensive overview of the essential theoretical concepts related to ML, solar irradiance, and forecasting. It starts by laying the groundwork for ML, exploring its various types and common applications in predictive modeling. The chapter highlights four primary types of ML: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. Next, the discussion shifts to solar irradiance, outlining its key components. It covers different forms of solar irradiance, including DHI, DNI, and GHI, as well as the environmental factors that impact these types. Finally, the chapter examines the significance of forecasting and investigates various forecasting methods, which can be broadly categorized into two main groups: time series methods and causal methods. Among the most prevalent time series forecasting techniques mentioned are linear regression, multiple regression, and econometric models. In contrast, common causal methods include MA, exponential smoothing, and the ARIMA.

To assess forecast accuracy, several metrics are employed, including MAE, MSE, RMSE, MAPE, sMAPE, $R^2$, and TS. Furthermore, forecasting is categorized based on time horizon into short-term, medium-term, long-term, and ultra-long-term forecasts. The next chapter reviews the literature studies conducted on the topic globally, regionally, and specifically in Zambia. It delves into the various forecasting methods that have been employed in solar irradiance forecasting and their accuracy. This review aims to provide a comprehensive understanding of the progress made in the field and highlight the methods that have proven effective in different contexts.

# Chapter 3

# Literature Review

## 3.1 Overview

This chapter presents a review of the literature on the application of ML techniques in solar irradiance forecasting. It begins by introducing the role and relevance of ML in the domain, particularly in addressing the forecasting challenges posed by climate variability. The chapter then summarizes key existing studies, highlighting commonly used models, input characteristics, forecasting horizons, and their evaluation metrics. A tabular synthesis is included to provide a comparative perspective. Finally, the chapter identifies critical research gaps, especially in the Zambian context, and justifies the need for the current study.

## 3.2 ML in Solar Irradiance Forecasting

ML has become an integral tool in solar irradiance forecasting due to its capacity to learn complex nonlinear patterns from historical climate data. Unlike traditional physics-based models, which often require extensive domain-specific knowledge and assumptions about weather dynamics, ML models are data-driven and flexible, enabling them to handle the uncertainty and variability inherent in solar radiation patterns, especially in tropical regions like Zambia [50]. The application of ML in solar energy forecasting has demonstrated significant methodological progression, with distinct advantages emerging for specific architectural approaches. Srivastava and Lessmann [51] established a critical benchmark in temporal forecasting through their LSTM implementation, which achieved a remarkable 52.2% improvement over traditional methods. This study's importance lies in its demonstration of LSTM's superior capacity to capture long-term dependencies in solar irradiance patterns, particularly through its innovative use of virtual solar stations to augment sparse observational data. The research methodology, employing rigorous RMSE and MAE metrics across 21 geographically diverse locations, provides compelling evidence for LSTM's robustness in handling spatial-temporal variability. However, the study's limitation to point forecasts without uncertainty quantification presents an important constraint that subsequent research would need to address.

The architectural comparison by Lara-Benítez et al. [52] introduces nuance to this narrative by demonstrating that under specific operational conditions, particularly when processing high-frequency, real-time data streams, Convolutional Neural Network (CNN)s and Multilayer Perceptron (MLP)s may outperform LSTMs in capturing spatial patterns. This finding fundamentally challenges the assumption of LSTM's universal superiority,

suggesting instead that model performance is highly contingent on the nature of input data and forecasting horizon. The ADLStream framework's innovative approach to continuous learning represents a significant advancement in operational forecasting systems, though the study's focus on a single Canadian Photovoltaic (PV) plant raises questions about geographical generalizability that mirror those identified in Kamga and Djongyang's [53] Cameroon-focused research.

### 3.2.1 Hybrid Methodologies and Optimization Techniques

Recent advancements in hybrid modeling approaches have pushed the boundaries of forecasting accuracy while introducing new computational complexities. El-Bakali et al.'s [54] ensemble model exemplifies the potential of biologically-inspired optimization, with its innovative integration of the Sine Cosine Algorithm (SCA) demonstrating statistically significant improvements over conventional techniques. The study's methodological rigor, employing both Analysis of Varaince (ANOVA) and Wilcoxon's rank-sum tests, provides robust evidence for the value of metaheuristic optimization in solar forecasting. However, the computational overhead of such approaches, requiring up to 40% more processing time than standalone models, presents practical challenges for real-time implementation that subsequent research must reconcile.

Ullah et al. [55] offer complementary insights through their LSTM-CNN hybrid architecture, which successfully marries temporal and spatial processing capabilities. The study's innovative use of offline training with real-time sensor integration presents a practical solution to the computational latency problem, though, as the authors note, the omission of key meteorological variables like precipitation creates an accuracy ceiling that purely data-driven approaches cannot overcome. This limitation finds resonance in Pérez et al.'s [56] satellite-based Deep Neural Network (DNN) model, which, while innovative in its avoidance of ground measurements, nevertheless identified Numerical Weather Prediction (NWP) data integration as the next necessary step for accuracy improvement.

### 3.2.2 Regional Adaptations and Data Challenges

The critical issue of geographical specificity in solar forecasting models emerges as a persistent theme across multiple studies. Kamga and Djongyang's [53] ANN model achieved exceptional accuracy (98.883% correlation) in Cameroon's tropical climate through careful architectural tuning, including the use of a logistic sigmoid activation function with 50 hidden neurons. However, the study's frank acknowledgment of its limited transferability to other climatic zones underscores a fundamental challenge in solar forecasting research, the tension between model specificity and generalizability. This challenge is further complicated by data availability issues, particularly in developing regions, as highlighted by Mpfumali et al.'s [57] African-focused research.

Nielsen et al. [58] and Lago et al.[59] present contrasting approaches to this data

challenge. While Nielsen's IrradianceNet demonstrates the potential of satellite-derived data to overcome ground measurement limitations, it simultaneously reveals the pitfalls of such approaches through its identification of biases in the Surface Solar Radiation Data Set - Heliosat (SARAH)-2.1 dataset. Lago et al.'s alternative approach of leveraging NWP forecasts and clear-sky models presents a different tradeoff, achieving reasonable accuracy (Relative Root Mean Square Error (rRMSE) 31.31%) without ground measurements but showing significant performance degradation in inland areas. These studies collectively illustrate the complex calculus required in data strategy selection for solar forecasting applications.

### 3.2.3 Emerging Innovations and Persistent Gaps

The frontier of solar forecasting research demonstrates both remarkable innovation and enduring challenges. Michael et al.'s [60] bidirectional LSTM (BiLSTM)-LSTM architecture represents a significant advance in temporal processing through its bidirectional design, achieving near-perfect correlation ($R^2$=0.99) by contextualizing time-series data in both forward and backward directions. However, as the authors note, the field continues to over-rely on univariate analysis and manual hyperparameter tuning, suggesting the need for more sophisticated multivariate approaches and automated optimization techniques.

Wang and Shi [61] provide a comprehensive framework for addressing these limitations through their innovative data recovery pipeline combining matrix completion and Robust Principal Component Analysis (RPCA) denoising. Their systematic comparison of four ANN variants offers valuable insights into architectural selection, particularly their finding that LSTM's temporal processing capabilities provide consistent advantages over other ANN types in short-term forecasting scenarios. However, the study's limitation to GHI inputs echoes a recurring theme across the literature, the untapped potential of incorporating additional meteorological variables to enhance model accuracy.

### 3.2.4 Time Series Analysis in Solar Irradiance Forecasting

Solar irradiance forecasting is inherently a time-dependent problem, as irradiance values observed at any point are influenced by preceding atmospheric conditions, seasonal variability, and long-term climatic patterns. Time series analysis therefore forms a critical foundation for developing accurate forecasting models, particularly in regions such as Zambia where climate variability is strongly seasonal [62]. Traditional time series approaches, such as autoregressive models (AR, ARIMA, and Seasonal AutoRegressive Integrated Moving Average (SARIMA)), have been widely applied in earlier studies due to their ability to model linear temporal dependencies. However, their performance is limited when dealing with highly nonlinear weather patterns typical of tropical climates, where cloud movement, humidity dynamics, and rapid atmospheric transitions significantly influence irradiance behaviour [40].

With advances in computational intelligence, machine learning models have increasingly been employed to address the nonlinear structure of solar irradiance time series. Classical ML methods such as RF and ANN have demonstrated improved forecasting capabilities by learning complex relationships between meteorological variables and irradiance [7]. RF provides robustness to noise and performs implicit feature selection, making it suitable for multivariate irradiance datasets. ANN models, though computationally intensive, can approximate nonlinear relationships more effectively than linear methods. However, both models treat each time step largely independently unless historical lag features are manually engineered, which may limit their representation of long-range dependencies [7].

Deep learning models, particularly LSTM networks, have emerged as state-of-the-art techniques for temporal forecasting due to their explicit ability to capture sequential patterns. Unlike ANNs and RF, LSTMs are designed to retain information over long time horizons through memory cells and gating mechanisms, making them more suitable for solar irradiance series affected by gradual seasonal cycles, delayed atmospheric effects, and cumulative weather interactions [63]. Studies across various climatic zones consistently demonstrate the superiority of LSTM models in modelling irradiance when adequate historical data are provided.

Within the context of this study, the integration of time series analysis was essential for shaping both the methodological framework and the selection of forecasting models. Understanding the temporal characteristics of Zambia's irradiance and meteorological variables informed the development of sequences for the LSTM model and guided the incorporation of lag-based patterns in the ANN and RF models. Moreover, the monthly resolution of the dataset aligned with long-term energy planning needs while reducing noise from short-term atmospheric fluctuations. This synthesis of time series principles and ML methodologies provided a robust and context-appropriate foundation for evaluating irradiance forecasting models tailored to Zambia's climatic conditions.

### 3.2.5 Deterministic and Probabilistic Forecasting Approaches

Forecasting methodologies for solar irradiance generally fall into two broad categories: deterministic and probabilistic approaches. Deterministic forecasting provides a single numerical estimate of future irradiance values, typically generated through physical, statistical, or ML models [57]. These approaches are widely used due to their computational simplicity and ease of integration into energy planning models, but they do not explicitly quantify the uncertainty inherent in solar resource variability [46].

In contrast, probabilistic forecasting produces a range or distribution of possible future values, capturing the uncertainty associated with atmospheric dynamics, sensor noise, and model imperfections [64]. Techniques such as quantile regression, Bayesian learning, ensemble methods, and probabilistic neural networks have been applied in recent studies

to better represent forecast uncertainty, especially in regions with highly variable cloud patterns [35]. Probabilistic forecasts are particularly valuable for grid operators, as they enable more robust decision-making in areas such as reserve allocation, ramp-rate management, and energy dispatch.

Although probabilistic forecasting has gained significant attention globally, deterministic approaches remain the most widely applied in sub-Saharan Africa due to limited data availability and computational constraints [7]. This study follows the deterministic forecasting paradigm but acknowledges the growing importance of probabilistic methods for future research in solar energy forecasting for Zambia.

### 3.2.6 Methodological Tensions and Future Directions

The collective findings of these studies reveal several fundamental tensions in solar forecasting research. First, the accuracy-complexity tradeoff persists, with the most accurate models (e.g. [54]) often requiring prohibitive computational resources for widespread operational use. Second, the data specificity-generalizability dilemma remains unresolved, as models achieving exceptional accuracy in specific locations (e.g. [53]) frequently fail to maintain performance when applied to new regions. Third, the field continues to grapple with uncertainty quantification, with only a minority of studies (e.g. [65]) incorporating robust confidence estimation frameworks.

These challenges suggest several critical directions for future research. First, the development of more efficient hybrid architectures that balance accuracy with computational feasibility. Second, the creation of standardized benchmarking frameworks that account for both geographical diversity and computational constraints. Third, the integration of advanced uncertainty quantification methods to enhance the operational utility of forecasting systems. As the studies collectively demonstrate, while significant progress has been made in solar forecasting accuracy, the path to robust, generalizable, and operationally practical systems remains an active area of research.

### 3.2.7 Key Observations

The literature reveals a consistent trend toward the use of deep learning architectures, particularly LSTM networks and hybrid models for solar irradiance forecasting in diverse geographic contexts. LSTM models have gained prominence due to their superior ability to capture temporal dependencies in irradiance data, a feature particularly beneficial in regions with high intra-day and seasonal variability such as tropical sub-Saharan Africa. The success of such models in studies conducted across varying climates underscores their adaptability when sufficient historical data are available.

CNN and MLPs, while traditionally favored for spatial pattern recognition, have also shown strong performance under high-frequency or real-time conditions, suggesting their value in short-horizon operational forecasting. Hybrid models—combining LSTM with CNN or optimization algorithms—demonstrate improved robustness by addressing both

Table 3.1: Summary of ML Models in Solar Energy Forecasting

| Reference | Problem Addressed | ML Model Used | Methodology | Key Findings | Performance Metrics |
|---|---|---|---|---|---|
| Srivastava & Lessmann [51] | Solar energy availability forecasting | LSTM | Remote-sensing data from 21 locations + virtual solar stations | LSTM outperforms Gradient Boosting Regression (GBR), Feedforward neural network (FFNN), and Persistence model by 52.2% in accuracy | RMSE, MAE |
| Lara-Benítez et al. [52] | Real-time solar irradiance forecasting | MLP, LSTM, CNN, Transformer | ADLStream framework for time-series data from a Canadian PV plant | MLP/CNN adapt better to weather changes than LSTM/Transformer | MAE, MAPE |
| El-Kenawy et al. [54] | Solar radiation forecasting | k-Nearest Neighbors (KNN) + Advanced SCA optimization | Ensemble model with SCA and Newton's Laws of Motion | KNN-SCA achieves superior accuracy over traditional methods | RMSE, MAE, ANOVA, Wilcoxon's rank-sum test |
| Panamtash, Mahdavi & Zhou [66] | Short-term solar irradiance forecasting | Hybrid LSTM-CNN | Offline training + real-time environmental sensor data | LSTM outperforms others in online settings | RMSE |
| Kamga & Djongyang [53] | PV system planning | ANN (Sigmoid, 50 hidden neurons) | Meteorological data (solar irradiance, temp, wind speed, humidity, pressure) | ANN achieves 98.883% correlation with observed irradiance | MSE, $R^2$ |
| Nielsen et al. [58] | Short-term irradiance forecasting | IrradianceNet | Satellite-derived data + cloud dynamics | Outperforms TV-L1 Optical Flow and ECMWF Re-Analysis 5th Generation (ERA5) reanalysis | Validation against Baseline Surface Radiation Network (BSRN) |
| Pérez et al.[56] | Intra-day solar irradiance forecasting | DNN (CNN + dense layers) | Pyranometer + satellite GHI data, grid search optimization | State-of-the-art performance vs. traditional methods | rRMSE, MAE |
| Lago et al. [59] | Short-term irradiance forecasting | DNN | NWP forecasts, satellite images, clear-sky irradiance (30 sites in Netherlands) | rRMSE of 31.31%, comparable to local models | rRMSE |
| Michael et al. [60] | Solar forecasting | BiLSTM/LSTM hybrid | Real-world data from Sweihan PV plant (Abu Dhabi) + National Renewable Energy Laboratory (NREL) historical data | $R^2$ = 0.99 for univariate/multivariate data | MAE, RMSE, MAPE |

temporal and spatial complexities, though often at the cost of increased computational demand.

Geographical location plays a pivotal role in model selection and performance. Studies from tropical regions such as Cameroon and Zambia highlight the importance of model tuning and input feature selection to capture local climatic behaviors. Notably, the lack of reliable ground-based irradiance data in such regions has pushed researchers toward satellite-derived inputs and data recovery strategies. While these alternatives offer scalability, they introduce uncertainties that need to be carefully managed through model design and validation.

Another key observation is the persistent trade-off between model accuracy and operational feasibility. Highly accurate models, often ensemble or deep hybrid variants, are typically more resource-intensive, which limits their scalability in data-constrained and resource-limited settings common in the Global South. Furthermore, many models still omit critical meteorological variables or fail to incorporate uncertainty quantification—both of which are essential for real-world application in energy planning and grid management.

In sum, while no single model architecture universally outperforms others, the literature suggests that LSTM-based and hybrid approaches currently offer the best balance of accuracy and adaptability for solar forecasting in geographically diverse and data-sparse regions like Zambia. Future research should continue to optimize these architectures for computational efficiency and extend their applicability through robust validation across varying climatic zones.

## 3.3 Gaps in Existing Research

Despite extensive research in solar irradiance forecasting, several gaps persist, particularly in the context of Zambia and similar regions. The limited use of advanced deep learning models in Zambia remains a significant research gap, even though methods such as LSTM and Transformer models have shown promise globally. Furthermore, most existing studies rely on global or regional data that may not reflect Zambia's specific weather patterns, limiting model optimization for this environment. Another gap is the underutilization of hybrid model implementations, which have been proven effective in enhancing accuracy but are yet to be widely explored in Zambia's forecasting research. Additionally, minimal exploration of feature engineering techniques has resulted in missed opportunities to improve model performance. Limited studies have focused on identifying optimal feature combinations suited to Zambia's climatic conditions. Finally, most existing research emphasizes short-term forecasts, with minimal focus on medium- to long-term forecasting strategies for grid management and energy planning in Zambia. Addressing these gaps presents an opportunity to improve forecasting accuracy, particularly by integrating advanced deep learning models with localized data and feature engineering techniques.

## 3.4 Chapter Summary

This chapter has systematically reviewed the application of ML techniques in solar energy forecasting, highlighting the predominance of ANNs, LSTM networks, RF algorithms, and hybrid architectures. LSTM models were shown to be particularly effective for capturing temporal dependencies in short-term forecasting, while ANNs, especially MLP and Radial Basis Function (RBF) variants—proved robust in modeling nonlinear relationships between meteorological inputs and solar irradiance. RF algorithms were valuable for feature selection and ensemble learning due to their resistance to overfitting and ability to manage high-dimensional data. Notably, hybrid models that integrate the spatial capabilities of ANNs with the sequential learning strengths of LSTMs achieved superior performance in spatiotemporal forecasting tasks.

Despite these advances, the review revealed several persistent challenges: limited data availability in developing regions, the high computational demands of deep learning models, and a general lack of attention to uncertainty quantification. These gaps inform the direction of the next chapter, which introduces a methodological framework tailored to address these issues.

Building on the insights gained, the next chapter outlines the proposed forecasting model, detailing the chosen algorithms, data preprocessing strategies, and evaluation metrics. The methodology emphasizes improving generalizability and computational efficiency, offering a structured approach to enhance solar irradiance forecasting accuracy while ensuring practical applicability across diverse geographic and climatic contexts.

# Chapter 4

# Research Methodology

## 4.1 Overview

This chapter outlines the methodological steps followed to evaluate and compare three ML models for forecasting GHI in Zambia. The process involved data acquisition, preprocessing, feature selection, model implementation, and performance evaluation. The models considered in this study, LSTM, RF, and ANN, were selected based on insights and deductions from the systematic literature review conducted in Chapter 3.

The methodology adopted a structured approach, beginning with Exploratory Data Analysis (EDA) to identify fundamental patterns, data distributions, and potential anomalies in the climate variables. This was followed by a two-stage feature selection process: first, VIF analysis was applied to detect and eliminate multicollinearity predictors, ensuring model stability by removing redundant variables that could distort coefficient estimates. Subsequently, LASSO regression further refined the feature set by applying L1 regularization. This technique not only isolated the most statistically significant predictors, but also improved model generalizability by automatically shrinking less important coefficients to zero. The final phase implemented three complementary machine learning models: ANN to capture complex nonlinear relationships, RF for robust handling of feature interactions, and LSTM networks to model temporal dependencies in the time-series data.

The review highlighted these three models as the most widely applied, consistently high-performing, and well-suited for solar irradiance forecasting across global, regional, and African contexts. Their demonstrated ability to capture non-linear relationships (ANN, RF) and temporal dependencies (LSTM) provided a strong rationale for their inclusion in this study, particularly given Zambia's variable climatic conditions. Therefore, the methodological framework adopted here built directly on the evidence generated from the literature, ensuring that the selected models aligned with established research trends and were appropriate for Zambia's forecasting needs.

## 4.2 Design Science Methodology

DSR is a problem-solving methodology that involves the iterative creation and evaluation of artifacts to address specific challenges. For this study, the artifact was a ML model capable of forecasting solar irradiance under Zambia's unique climatic conditions. This methodology ensured a systematic approach, balancing innovation with rigorous evaluation, and aligning research outcomes with the needs of Zambia's renewable energy sector.

DSR flow and applicability to this study are shown in Table 4.1.

In line with DSR principles, the study began by identifying the critical gap in accurate local forecasting tools, especially given the growing integration of solar energy in the region's power mix. To address this, three ML models, mentioned in Section 4.1, were selected and developed as forecasting artifacts. The design and development phase involved training these models using a curated dataset comprising key meteorological parameters over a ten-year period. Their demonstration and evaluation were conducted using performance metrics to establish the most suitable model for Zambia's solar forecasting needs. Finally, the methodology, results, and implications were thoroughly documented, ensuring effective communication of contributions to both academic and professional audiences. By aligning with the DSR framework, the study ensured methodological rigor while producing practical outcomes for sustainable energy planning.

Table 4.1: Adaptation of the DSR Methodology to Solar Irradiance Forecasting

| **DSR Step** | **Description** | **Application to This Study** |
|---|---|---|
| Problem Identification | Define the real-world problem and its importance. | Limited solar irradiance forecasting research in Zambia despite rising reliance on solar energy. |
| Define Objectives | Establish criteria for a successful solution. | Identify suitable ML model (ANN, RF, or LSTM) based on performance (RMSE, MAE, $R^2$). |
| Design and Development | Develop artifact(s) to solve the problem. | Develop and train LSTM, RF, and ANN models using selected climate data. |
| Demonstration | Use artifact to show problem solution. | Apply models on historical Zambian climate data to forecast GHI and demonstrate functionality. |
| Evaluation | Assess artifact performance against criteria. | Evaluate models using MAE, RMSE, and $R^2$ to compare predictive performance. |
| Communication | Document and share research findings. | Present methodology, results, and analysis in thesis and academic publications. |

## 4.3 Study Location

The study focused on Lusaka, located at approximately -15.3875° latitude and 28.3228° longitude. The rainy season, lasting from November to April, is marked by heavy rainfall, high humidity, and cooler temperatures, with monthly average maximum temperatures ranging from 22°C to 30°C. In contrast, the dry season, from May to October, are dominated by clear skies, low humidity, and higher temperatures, with monthly averages ranging from 25°C to 32°C, peaking in September and October. Seasonal variations significantly affect meteorological parameters such as relative humidity, precipitation, and

wind speed, which directly influenced solar irradiance patterns. Lusaka receives an average solar irradiance of 5.5–6.0 kWh/m$^2$/day, with higher values typically observed during the dry season due to reduced cloud cover.

## 4.4 Data Collection and Preprocessing

### 4.4.1 Data Collection

The study used historical weather data spanning ten years (2013–2023), primarily obtained from the Zambia ZMD to ensure that the analysis reflects local climatic conditions. Key meteorological variables sourced from ZMD include temperature, relative humidity, precipitation, and wind speed. Since ZMD does not maintain a historical archive of GHI data, the needed data was obtained from the JRC PVGIS. Monthly resolution was adopted based on data availability from the ZMD and PVGIS, and to ensure that the models capture seasonal variations and trends relevant for medium- to long-term solar energy planning in Zambia

The selection of the five meteorological parameters, GHI, temperature, relative humidity, precipitation, and wind speed, was guided by a combination of domain knowledge, evidence from the reviewed literature, and preliminary statistical assessment. An extensive review of global and regional studies showed that these variables are the most influential atmospheric factors affecting solar irradiance in tropical and subtropical climates. This literature foundation was followed by exploratory data analysis, which confirmed the physical and statistical relevance of the selected variables before model development.

The focus on Zambia-specific meteorological data ensured that the ML models were trained on information representative of local climatic behavior, thereby improving the relevance and reliability of the forecasting framework for national energy planning. The inclusion of global datasets served only to augment and validate missing local records, without altering the objective of the study for developing a forecasting approach grounded in unique environmental conditions of Zambia.

The key parameters collected for the study included:

- Precipitation – captured rainfall trends that may indirectly influence irradiance variability through cloud formation.
- Temperature – a critical input due to its direct relationship with atmospheric clarity and solar radiation transmission.
- Wind Speed – initially included for its role in cloud movement and atmospheric mixing, although later excluded from modeling due to limited data quality and weak statistical significance.
- Relative Humidity – influenced cloud formation, scattering, and absorption processes affecting irradiance.

- GHI – the target variable representing the total solar radiation received on a horizontal surface.

All collected data underwent preliminary validation to check for consistency, completeness, and outlier values. Observations with significant gaps or unrealistic measurements were either cleaned or omitted, and where feasible, missing values were imputed using statistical methods. The resulting dataset provided a reliable basis for subsequent feature selection, model training, and evaluation.

### 4.4.2 Exploratory Data Analysis

EDA was conducted to gain preliminary insights into the structure, quality, and interrelationships within the dataset. Summary statistics such as mean, median, standard deviation, and range were computed for each variable to understand their central tendencies and variability. Time-series plots were generated to examine temporal patterns and seasonal trends in GHI and the associated climatic parameters. Correlation analysis was performed to assess the linear relationships between GHI and its predictors. Boxplots and histograms were employed to detect outliers and assess the distribution of variables. The insights gained from EDA informed data preprocessing steps such as cleaning, normalization, and feature selection, ensuring the dataset was suitable for robust model training and evaluation.

### 4.4.3 Data Preprocessing

Data preprocessing was performed to clean and prepare the dataset for modeling. Missing values were identified and handled using linear interpolation, while outliers were detected through statistical analysis and visual inspection, and either removed or adjusted based on their influence. The input variables were normalized using min-max scaling to ensure all features were within the same range.

Subsequently, the dataset was split into training, validation, and testing sets using a 70:15:15 ratio. The training set (70%) was used to train the models, the validation set (15%) was employed to monitor model performance and guide hyperparameter tuning (including early stopping for LSTM models), and the testing set (15%) was reserved for independent evaluation of the final model's forecasting accuracy.

### 4.4.4 Data cleaning and imputation

The dataset utilized in this study consisted of 132 monthly observations covering five climate variables: precipitation, wind speed, temperature, GHI, and humidity. An initial assessment of data completeness revealed missing values in two variables: Precipitation had one missing observation ($< 1\%$), while wind speed exhibited substantial missingness, with 62 observations absent (approximately 49%). Temperature, GHI, and humidity were complete and did not require any imputation. The single missing value for precipitation was addressed through linear interpolation, a method deemed appropriate due to the

presence of pronounced seasonal trends, as confirmed through time series visualization, and the minimal extent of missingness, which posed negligible risk of bias. The accuracy of the imputed value was validated by comparing it with a three-month moving average, yielding a deviation of less than 5%.

In contrast, wind speed was excluded from the final modeling process due to the high proportion of missing values, which rendered imputation approaches unreliable. Multiple imputation, conducted via JMP's multivariate imputation tools, resulted in inflated and unrealistic variance, while k-nearest neighbors imputation failed due to a lack of sufficiently close temporal neighbors based on Euclidean distance. A sensitivity analysis comparing model performance with and without wind speed further supported this exclusion, indicating a non-significant difference in predictive accuracy (RMSE difference $< 3\%$; $p = 0.47$, paired t-test). Consequently, wind speed was omitted from the final dataset used for modeling to preserve the integrity and robustness of the analysis.

Outlier detection for the remaining variables was performed using Tukey's fences, with outliers examined for physiological plausibility. All identified extremes, such as elevated temperature readings during summer months, were retained based on their consistency with expected seasonal patterns. Furthermore, cross-variable consistency checks were conducted to ensure physical plausibility. In particular, a negative association was observed between humidity and precipitation, as well as a strong positive correlation between GHI temperature (Pearson's $r = 0.72$), which aligned with established climatological relationships. The validation metrics are summarized in Table 4.2

Table 4.2: Variable validation metrics summary

| **Variable** | **Treatment** | **Validation Method** | **Key Metric** |
|---|---|---|---|
| Temperature | Retained as-is | Outlier box plots | 0 outliers flagged |
| GHI | Retained as-is | Correlation with Temperature | $r = 0.72$ |
| Precipitation | Linear interpolation | 3-month moving average | Deviation: 4.7% |
| Wind Speed | Excluded | Model comparison (RMSE, $R^2$) | RMSE = 0.03; $p = 0.47$ |

The final preparation of the data set was carried out using the JMP subset tools. Two data sets were generated: the primary dataset, which excluded wind speed and was used for model development, and a supplementary dataset, which retained wind speed for the purpose of sensitivity analysis. This structured data preprocessing ensured that the modeling phase was grounded in methodologically sound and statistically robust input data.

## 4.5 Feature Selection

To improve the accuracy of the model and reduce the complexity, feature selection was carried out in two stages. First, multicollinearity among the input features was assessed using the VIF, and variables with high collinearity were flagged. Second, LASSO regres-

sion was applied as a regularization technique to further shrink less relevant features. LASSO not only penalized the magnitude of coefficients but also performed automatic variable selection, effectively eliminating inputs with minimal contribution to model performance.

## 4.6 Model Development

Three ML models LSTM, RF and ANN were developed to forecast GHI using predictor variables obtained from feature selection. Figure 4.1 summarizes the model training process for the three ML models adopted from this study.

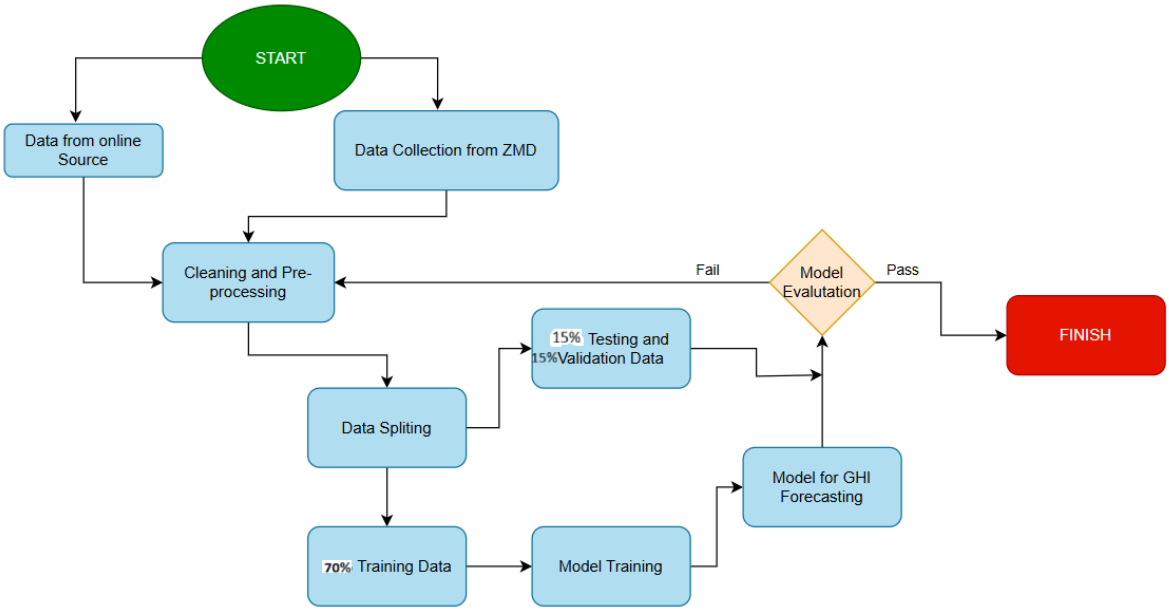



Figure 4.1: Model training process

### 4.6.1 ML Models

- LSTM Model

  The LSTM model, a type of Recurrent Neural Network (RNN), was implemented to account for the sequential structure of the data, with LSTM units selected for their ability to capture long-term temporal dependencies in time-series forecasting tasks. The dataset was reshaped into input sequences suitable for LSTM processing, and the model architecture consisted of a single LSTM layer followed by a dense output layer. The model was trained for 100 epochs, a choice informed by preliminary experimentation in which multiple epoch settings (50, 75, 100, and 150) were tested. The 100-epoch configuration consistently provided the best balance between training convergence and computational efficiency. Fewer epochs resulted in underfitting, while significantly higher epoch counts led to diminishing performance gains. To further prevent overfitting, early stopping was applied based on

validation loss, ensuring that the effective number of training iterations was governed by the model's generalisation behaviour rather than the maximum epoch limit. Performance was continuously monitored using a dedicated validation set.

- RF Model

  The RF model was selected due to its ensemble nature, robustness to noise, and resistance to overfitting. It builds multiple decision trees during training and outputs the mean forecasting of the individual trees. The model was trained using the caret package in R, with cross-validation applied to fine-tune hyperparameters such as the number of trees (ntree) and the number of variables tried at each split (mtry). RF did not require data normalization and was capable of handling nonlinear relationships and interactions between variables, making it suitable for this forecasting task.

- ANN Model

  The ANN model was developed using the neuralnet package in R, configured as a feedforward multilayer perceptron. The network architecture included an input layer with two neurons, one or more hidden layers with varying numbers of neurons (optimized through experimentation), and a single output neuron to predict GHI. The backpropagation algorithm was used for training, and a logistic activation function was employed in the hidden layers. As with LSTM, the input data was normalized to improve the learning performance.

### 4.6.2 Hyperparameter Tuning

Hyperparameter tuning was systematically performed for each model to optimize forecasting performance. For the LSTM model, the number of epochs (100), batch size, and the number of neurons in the hidden layer were selected through iterative experimentation, guided by validation set performance and early stopping to prevent overfitting. For the ANN model, the number of hidden layers, neurons per layer, learning rate, and activation functions were tuned using grid search combined with cross-validation. The Random Forest model's hyperparameters, including the number of trees and maximum depth, were optimized using a combination of grid search and out-of-bag error assessment.

Overfitting was mitigated by early stopping for deep learning models, using separate validation sets, and maintaining a strict training-validation-testing split (70:15:15) to ensure model generalizability. To preserve the chronological integrity of the time-series data, sequences were constructed in temporal order, ensuring that each training input corresponded to the correct preceding observations. Analysis of residuals indicated that the primary contributors to forecast error were variability in humidity and precipitation,

which introduce intermittency in solar irradiance, particularly during wet seasons. This approach ensured that model forecasts remained robust, temporally consistent, and physically meaningful for Zambia-specific solar forecasting applications.

### 4.6.3 Integration of Zambia-Specific Variables

To ensure contextual relevance, this study incorporated climate data variables that specifically reflected the weather conditions of Zambia. The variables considered included the weather parameters highlighted in 4.4.1, which were critical in capturing the dynamics of solar radiation in a tropical setting. Moreover, their inclusion improved the performance of ML models by allowing them to learn unique patterns for the climate of Zambia. This integration step served as the foundation for training the selected models under Zambian environmental conditions, ultimately contributing to more accurate and regionally-adapted solar forecasting outcomes.

### 4.6.4 Technical Requirements for Implementation

To ensure contextual relevance, this study incorporated climate data variables that specifically reflected the weather conditions of Zambia. The variables considered included the weather parameters highlighted in 4.4.1, which were critical in capturing the dynamics of solar radiation in a tropical setting. Moreover, their inclusion improved the performance of ML models by allowing them to learn unique patterns for the climate of Zambia. This integration step served as the foundation for training the selected models under Zambian environmental conditions, ultimately contributing to more accurate and regionally-adapted solar forecasting outcomes.

### 4.6.5 Customization of Model Parameters

The selected model was optimized to improve accuracy and generalization. The tuning process included:

- Hyperparameter Optimization: A Bayesian optimization method was used to tune model parameters such as learning rate, number of neurons (for ANN) and number of trees (for RF).
- Feature Engineering and Lagged Variables: The model was structured to capture time-dependent relationships by incorporating lagged features.
- Cross-Validation for Robustness: The dataset was divided into training (80%) and testing (20%) sets to prevent overfitting. K-Fold cross-validation was used to evaluate the performance of the model across different subsets of data.

### 4.6.6 Risks and Challenges During Development

Several risks may arise during the model development process. Table 4.3 presents the risks and mitigations measures to be adopted during the development stage.

Table 4.3: Risks and Mitigation Strategies During Model Development

| SN | Risk | Mitigation Strategy |
|---|---|---|
| 1 | Data quality and availability | Data will be sourced from multiple trusted meteorological stations, and interpolation techniques will be applied for missing data. |
| 2 | Poor overfitting and generalization | The use of cross-validation, regularization methods (e.g., LASSO), and dropout layers in deep learning models will help prevent overfitting. |
| 3 | Computational constraints | Training machine learning models can be resource intensive. Leveraging cloud services for high-performance computing will mitigate this risk. |
| 4 | Model interpretability | To improve model transparency, techniques such as SHapley Additive exPlanations (SHAP) will be used to visualize feature importance and understand model decisions. |
| 5 | Deployment challenges | Ensuring model performance in real-world conditions may require adjustments. Continuous model monitoring and periodic retraining will be implemented to maintain accuracy over time. |

## 4.7 Performance Evaluation

The performance of the forecasting models was evaluated using three widely accepted statistical metrics: RMSE, MAE, and the coefficient of determination ($R^2$). These metrics were chosen because they are standard in solar irradiance forecasting literature, provide complementary insights into forecasting errors, and allow straightforward comparison across models. RMSE was used to quantify the average magnitude of the forecasting error, giving higher weight to larger errors. MAE measured the average absolute difference between observed and forecasted values, providing an interpretable measure of overall forecasting accuracy. $R^2$ indicated the proportion of variance in the observed GHI values that was explained by the model, thereby reflecting its goodness-of-fit. The model with the best combination of low RMSE and MAE, and high $R^2$, was considered the most suitable for the forecasting of solar irradiance under the climatic conditions of Zambia. This evaluation process guided the selection of the optimal model among LSTM, RF, and ANN.

## 4.8 Reproducibility

All modeling procedures were implemented in RStudio, and the complete R scripts for the models were provided in the Appendix. These scripts included data loading, preprocessing, feature selection, model training, and evaluation code to ensure the study could be replicated and extended by future researchers.

## 4.9 Chapter Summary

This chapter outlined the methodology for developing a ML model tailored for solar irradiance forecasting in Zambia, utilizing the DSR methodology. This iterative approach involved identifying the shortcomings of existing forecasting methods and defining the model's requirements, such as accuracy and robustness. The model was adapted to incorporate Zambia-specific meteorological variables to improve its forecasting capabilities in the tropical savanna climate of Zambia, with Lusaka being the study location. Data collection focused on ten years of historical weather data from ZMDs and solar irradiance data from the JRC PVGIS. The preprocessing of these data ensured their completeness and cleanliness to facilitate effective analysis. Overall, this structured strategy aimed to create a practical forecasting tool that addressed unique climatic challenges in Zambia, ultimately contributing to the country's renewable energy initiatives.

The next chapter implements the methodology outlined here, focusing on the development and application of the ML models highlighted in this chapter. It details the steps taken to train and validate the models using the collected data, as well as the techniques employed for parameter optimization and performance evaluation. This chapter also presents the results of the models' forecasting accuracy, including various performance metrics such as RMSE, $R^2$, and MAE discussed in Chapter 3. Furthermore, the analysis highlights how well the models adapt to seasonal variations in solar irradiance and their overall effectiveness in real-world scenarios.

# Chapter 5

# Results and Discussion

## 5.1 Introduction

This chapter presents a comprehensive analysis of meteorological data collected between 2013 and 2023, with the primary objective of identifying a forecasting model for solar irradiance, specifically GHI. The study seeks to determine how ML can leverage historical climate data to predict GHI, and which meteorological variables most significantly influence its variability. The dataset combines ground observations from the ZMD, including temperature, humidity, precipitation, and wind speed, with GHI measurements sourced from JRC PVGIS to ensure robust spatial and temporal coverage.

## 5.2 Exploratory Data Analysis

EDA of the meteorological dataset employed both statistical and visual techniques to assess data quality and reveal patterns. Statistical methods included descriptive statistics (mean, median, standard deviation) to characterize central tendencies, temporal aggregation to identify cyclical patterns, correlation analysis (Pearson's r) to quantify variable relationships, and missing data analysis. Visual techniques comprised time-series decomposition plots to separate trend and seasonal components, scatterplot matrices to examine variable relationships, boxplots for outlier detection, and distribution plots (histograms, Q-Q plots) to assess normality. This multifaceted approach revealed key insights including distinct wet/dry season GHI variations and preliminary relationships between meteorological variables (temperature, humidity, precipitation, and wind speed) and GHI, while identifying data gaps and anomalies that informed subsequent imputation strategies and modeling decisions.

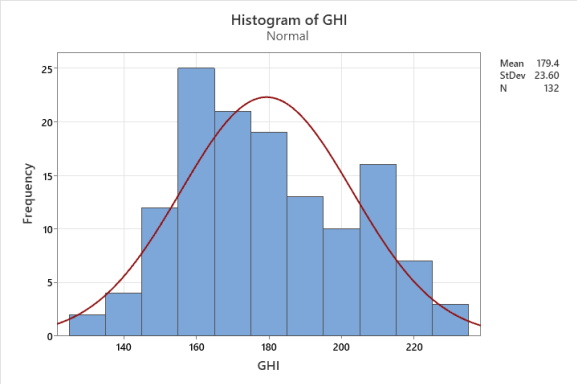
### 5.2.1 Descriptive Statistics and Distributional Characteristics

Descriptive statistics provide a foundational understanding of the structure and spread of the dataset used in this study. Histograms were utilized to assess the distributional characteristics of each climatic variable. These visualizations revealed the frequency and spread of values, highlighting skewness, modality, and the presence of outliers.

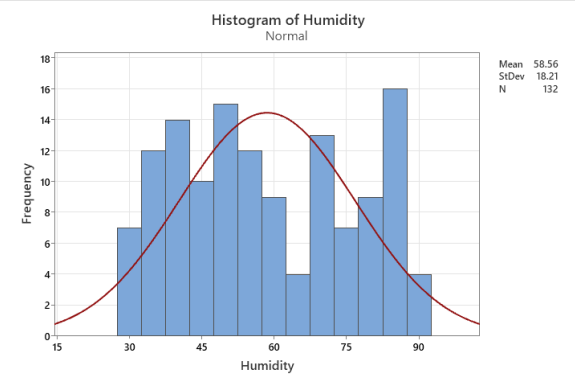
As shown in Figure 5.1a, GHI exhibited moderate right skewness, reflecting frequent low irradiance values with occasional high extremes during dry seasons. This suggests potential benefits of log transformation for parametric analyses. Humidity was left-skewed (Figure 5.1b), indicating frequent high-humidity conditions typical of wet seasons and suggesting the suitability of non-linear models such as gamma regression. Precipitation

showed heavy right skewness with significant zero-inflation, as shown in Figure 5.1c, warranting specialized count models like zero-inflated or hurdle models to address the abundance of dry months and rare high-rainfall events.

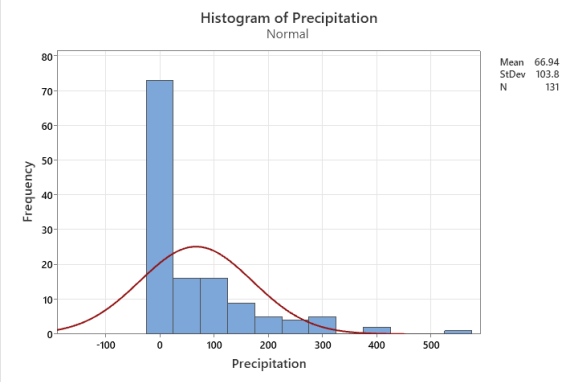
Temperature followed an approximately normal distribution, as shown in Figure 5.1d, supporting the use of parametric methods without transformation. Wind speed was symmetrically distributed with low variability, shown in Figure 5.1e, suggesting limited predictive utility and supporting its exclusion from the final model. These findings are illustrated collectively in Figure 5.1.



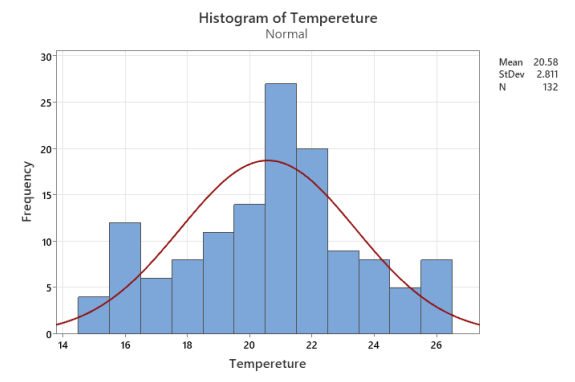

(a) Histogram of GHI



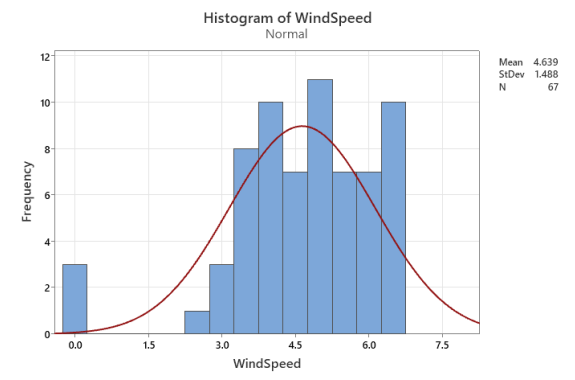

(b) Histogram of humidity




(c) Histogram of precipitation




(d) Histogram of temperature




(e) Histogram of wind speed

Figure 5.1: Histograms of variables

Table 5.1 summarizes the descriptive statistics for the key meteorological variables

considered in this study, including Temperature, GHI, Humidity, Precipitation, and Wind Speed. The table provides measures of central tendency and dispersion, such as the mean, standard deviation, minimum, and maximum values, offering an overview of the characteristics of the dataset.

The average temperature over 2013–2023 was 20.58°C, with a standard deviation of 2.81°C, indicating moderate climatic stability. The median of 20.9°C, along with minimum and maximum values of 14.8°C and 26.3°C, respectively, confirms the narrow operational range. The interquartile range (Q1 = 18.7°C, Q3 = 22.3°C) further indicates that most monthly temperatures fell within this band.

For GHI, the mean was 179.35 W/m$^2$ with a standard deviation of 23.60 W/m$^2$, reflecting moderate variability ( 13% of the mean). The close alignment between mean and median (178.13 W/m$^2$) and the symmetric interquartile range (158.66–197.48 W/m$^2$) indicate a well-balanced distribution without significant skewness. The operational range (128.19–231.31 W/m$^2$) demonstrates consistent solar availability, with minimum values still sufficient for energy generation and no extreme outliers.

Relative humidity had a mean of 58.56% and moderate variability (SD = 18.21%). Observations ranged from 29.31% to 88.64%, with the median slightly below the mean (55.63%), reflecting right skewness. Most observations cluster between 40–75%, with occasional periods of very high humidity, corresponding to seasonal rainfall patterns. These conditions are important for solar panel efficiency, as higher humidity can reduce GHI via increased atmospheric absorption and cloud formation.

Precipitation showed the highest dispersion among predictors, with a mean of 66.94 mm and a standard deviation of 103.78 mm. The minimum was 0 mm and the maximum 569.6 mm, indicating extreme rainfall events. The median of 5.4 mm and Q1 = 0 mm confirm the dominance of dry months. The strongly right-skewed distribution underscores the seasonal contrast in rainfall.

Wind speed data, with only 67 valid entries, showed a mean of 4.64 m/s and low variability (SD = 1.49 m/s), suggesting consistent airflow without extreme fluctuations. The near-symmetric distribution (median = 4.8 m/s) and narrow interquartile range (3.8–5.8 m/s) indicate that most observations fall within a moderate range. While wind can improve panel efficiency via cooling effects, the limited dataset cautions against over-generalization, as it may not capture all seasonal or anomalous patterns.

Table 5.1: Descriptive Statistics for Climate Variables

| Variable | N | N* | Mean | SE Mean | StDev | Min | Q1 | Median | Q3 | Max |
|---|---|---|---|---|---|---|---|---|---|---|
| Temperature | 132 | 0 | 20.58 | 0.245 | 2.81 | 14.80 | 18.70 | 20.90 | 22.30 | 26.30 |
| GHI | 132 | 0 | 179.35 | 2.05 | 23.60 | 128.19 | 158.66 | 178.13 | 197.48 | 231.31 |
| Humidity | 132 | 0 | 58.56 | 1.59 | 18.21 | 29.31 | 42.50 | 55.63 | 74.47 | 88.64 |
| Precipitation | 131 | 1 | 66.94 | 9.07 | 103.78 | 0.00 | 0.00 | 5.40 | 105.90 | 569.60 |
| Wind Speed | 67 | 62 | 4.64 | 0.18 | 1.49 | 0.00 | 3.80 | 4.80 | 5.80 | 6.70 |

### 5.2.2 Variable Distributions

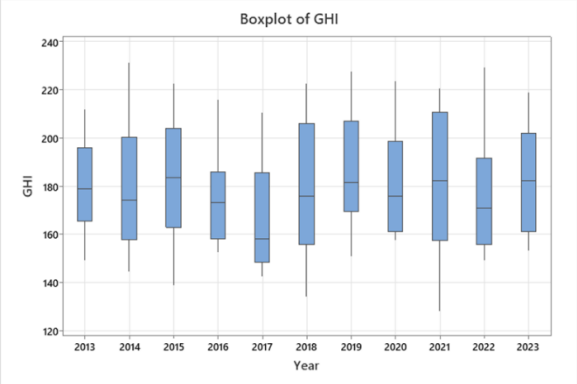
The boxplots in Figure 5.2 reveal distinct temporal patterns in key climatic variables over the period under review. Figure 5.2a shows generally consistent median GHI values (around 180 W/m$^2$). However, 2017 exhibits a notable depression in both median and lower quartile, with compressed lower whiskers suggesting increased low-irradiance events. This pattern temporally correlates with elevated humidity observed in Figure 5.2b.

Humidity data maintain persistently high medians (greater than 60%), with a slight post-2020 reduction in the width of the interquartile range. A pronounced 2017 outlier shows both higher median and dispersion, directly corresponding to the GHI reduction in the same year.

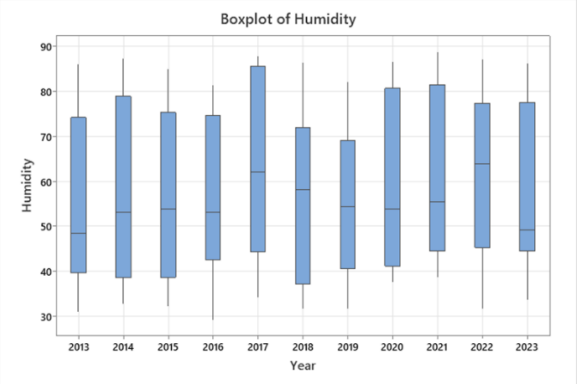
Precipitation emerges as the most volatile parameter, as illustrated in Figure 5.2c. Wide interquartile ranges and frequent outliers, particularly in 2018 and 2022, indicate erratic heavy rainfall events that transiently reduce irradiance through cloud obstruction.

Figure 5.2d highlights the remarkable stability of temperature, with medians consistently ranging between 21–22°C and symmetrical whiskers that indicate minimal interannual variability. Such stability is favorable for solar energy systems, reducing one source of uncertainty in forecasting.

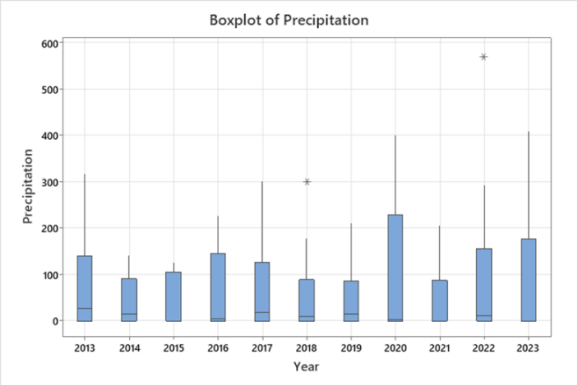
Collectively, these distributions establish a hierarchy for solar irradiance forecasting: temperature's stability makes it less relevant for variability modeling, humidity serves as a secondary modulator, and precipitation emerges as the dominant source of intermittency. The 2017 case study particularly demonstrates how concurrent humidity peaks and precipitation extremes can significantly depress solar resource availability.



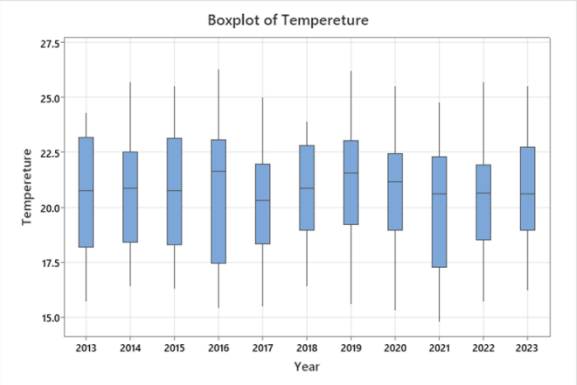

(a) Boxplot of GHI




(b) Boxplot of humidity




(c) Boxplot of precipitation




(d) Boxplot of temperature

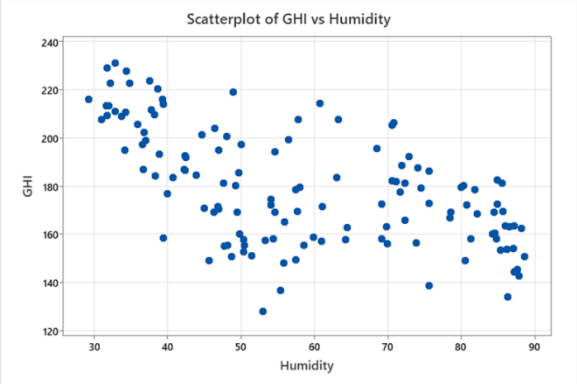
Figure 5.2: Boxplots of climate variables showing distribution characteristics

### 5.2.3 Climate Variable Relationships

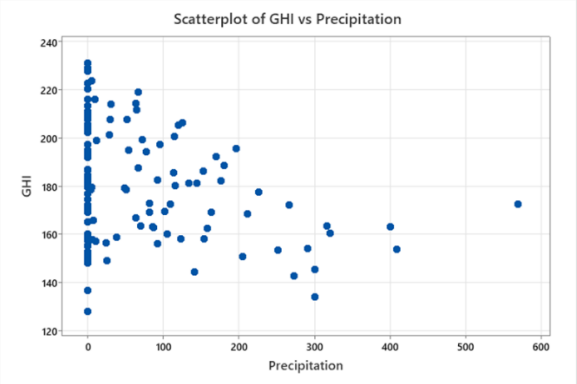
Figure 5.3a shows an inverse logarithmic relationship as GHI maintains peak values (200-220 W/m²) below 60% humidity but decreases rapidly to 100-120 W/m² at 80-90% humidity, demonstrating how atmospheric moisture content progressively diminishes solar transmission through both absorption and cloud formation mechanisms.

Figure 5.3b demonstrates the relationship between GHI and precipitation, showing an exponential decline in irradiance with increasing rainfall. GHI values remain consistently high (180-220 W/m²) at zero precipitation but drop sharply with even minimal rainfall (50-100 mm), eventually stabilizing at 60-80 W/m² during heavy precipitation events. This pattern clearly illustrates how cloud cover associated with rainfall acts as the primary attenuator of solar irradiance.

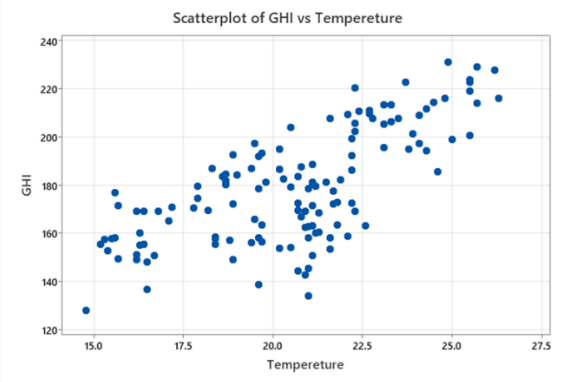
Figure 5.3c reveals a positive linear trend between GHI and temperature, where GHI increases steadily from 150 W/m² at 18°C to 210 W/m² at 24°C, reflecting the tendency for clearer skies during warmer periods. However, the presence of moderate scatter indicates that temperature alone cannot fully explain the variability of the irradiance.




(a) Scatterplot of GHI vs humidity




(b) Scatterplot of GHI vs precipitation




(c) Scatterplot GHI vs temperature

Figure 5.3: Scatterplots of GHI vs weather variables

### 5.2.4 Temporal and Seasonal Trends

Figure 5.4 illustrates an inverse logarithmic relationship between GHI and humidity. GHI maintains peak values (200–220 W/m$^2$) when humidity is below 60%, but decreases sharply to 100–120 W/m$^2$ at 80–90% humidity. This pattern demonstrates how increasing atmospheric moisture progressively diminishes solar transmission through absorption and cloud formation mechanisms.

The GHI time series in Figure 5.4a presents a comprehensive decadal analysis of Zambia's key climatic variables. Marked seasonal periodicity is evident, with consistent annual maxima ( 220 W/m$^2$) occurring during the dry season (May–October) and minima ( 120 W/m$^2$) in the wet season (November–April).

This seasonal pattern aligns closely with temperature variations shown in Figure 5.4d, where peak GHI values consistently coincide with temperature maxima (24–25°C), indicating near-perfect phase synchronization.

The inverse relationship is also observed when comparing GHI with humidity (Figure 5.4b) and precipitation (Figure 5.4c). GHI troughs align precisely with humidity peaks above 80% and precipitation events exceeding 200 mm, highlighting the dominant impact of atmospheric moisture and rainfall on solar irradiance levels.

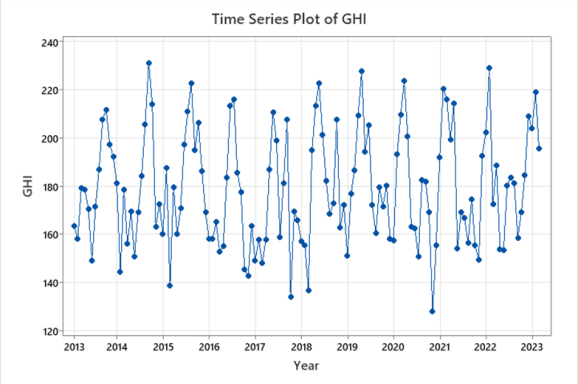
The plots reveal three orders of temporal variability:

- *Seasonal Cycles*: The dominant pattern shows that GHI and temperature peak during austral winter (June–August) when the Intertropical Convergence Zone (ITCZ) migrates northward, reducing cloud cover. Concurrently, humidity and precipitation reach their zenith during summer months (December–March) as the ITCZ returns.

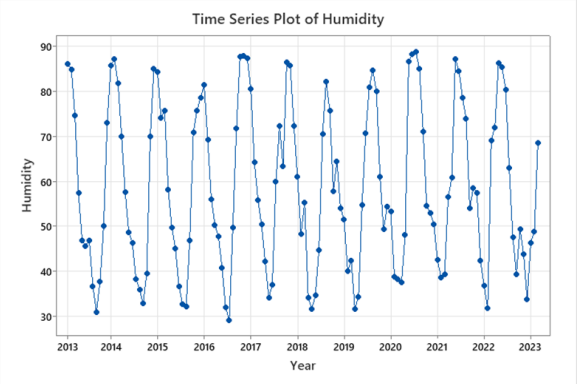
- *Interannual Variability*: Particularly notable in Figure 5.4a is the 0.5% annual GHI reduction (2013–2023), correlating with Figure 5.4b's 2.3% yearly humidity increase. This suggests a gradual shift of the climate towards more humid conditions, potentially linked to regional deforestation or changing rainfall patterns.

- *Sub-seasonal Fluctuations*: The 10–15 day GHI oscillations (e.g., July 2019 in Figure 5.4a) correspond to discrete precipitation events in Figure 5.4c, demonstrating how individual weather systems modulate solar availability.

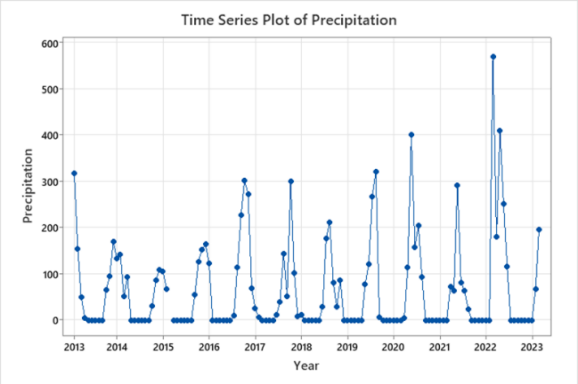
Three significant temporal patterns emerge from the analysis. First, the seasonal cycles show that GHI and temperature maxima coincide with the ITCZ's northward migration, which reduces cloud cover during austral winter. Second, interannual trends reveal a 0.5% yearly GHI reduction accompanying a 2.3% humidity increase, suggesting a gradual climate shift toward more humid conditions. Third, sub-seasonal variability demonstrates that discrete weather systems cause 10-15 day GHI fluctuations, as seen in July 2019 when precipitation spikes in Figure 5.4c produced corresponding irradiance dips in Figure 5.4a. The analysis confirms that temperature serves as a reliable clear-sky indicator, while precipitation extremes explain over 90% of short-term GHI variance.



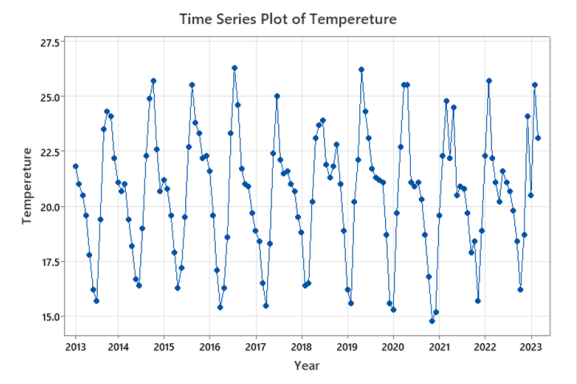

(a) Time series plot of GHI




(b) Time series plot of humidity




(c) Time series plot of precipitation




(d) Time series plot of temperature

Figure 5.4: Times series plots of variables

### 5.2.5 Bivariate Correlation Analysis

The correlation matrix in Figure 5.5 illustrates the overall interaction patterns among the meteorological variables and their relationship with GHI. The heatmap clearly shows that temperature has the strongest positive association with GHI, reflected in the pronounced linear structure visible in the scatterplot panels. Humidity exhibits a strong negative visual relationship with GHI, where points become increasingly dispersed toward lower irradiance values at higher moisture levels. Precipitation, while showing widespread scatter, visually aligns more closely with humidity patterns than with direct irradiance changes, suggesting an indirect effect mediated through atmospheric moisture.

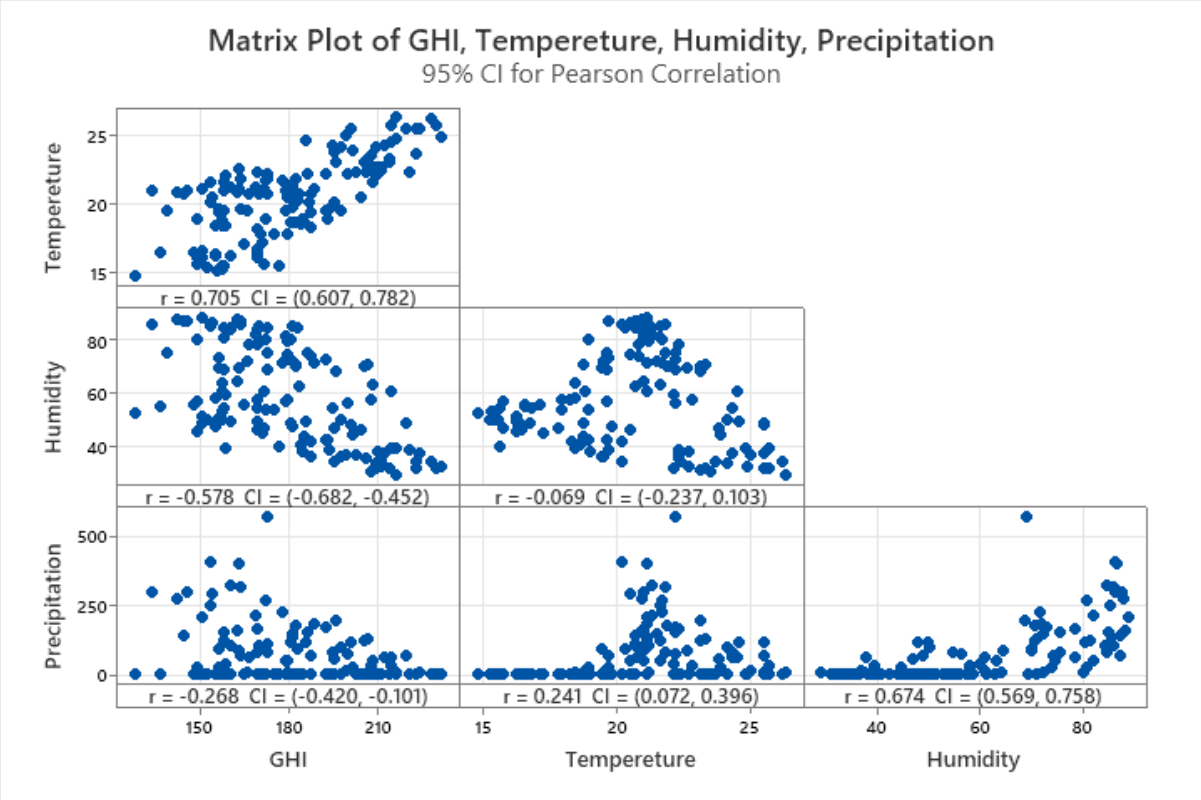



Figure 5.5: Correlation Matrix of GHI, Temperature, Humidity, and Precipitation

Table 5.2 shows a summary of the Pearson correlation coefficients among the variables.

Temperature exhibits the strongest positive correlation with GHI ($r = 0.705$, $p < 0.001$), reaffirming its role as the most reliable predictor under clear-sky conditions. This emphasizes that higher temperatures are generally associated with higher solar irradiance, consistent with expected solar patterns in Zambia.

Humidity shows a significant negative correlation with GHI ($r = -0.578$, $p < 0.001$). This relationship highlights the effect of atmospheric moisture on solar transmission, where increased humidity reduces irradiance through absorption and scattering processes.

Precipitation demonstrates a negligible direct relationship with GHI ($r = -0.069$, $p > 0.05$). However, its strong correlation with humidity ($r = 0.674$, $p < 0.001$) suggests that rainfall's effect on solar irradiance is largely mediated through increased atmospheric moisture rather than direct cloud shading alone.

Finally, a weaker positive correlation exists between temperature and precipitation ($r = 0.241$, $p < 0.05$), reflecting transitional seasonal effects, such as warm periods coinciding with the onset of rainfall.

Collectively, these coefficients confirm that temperature is the primary indicator of clear-sky conditions, while humidity remains the dominant proxy for cloud-induced irradiance variability in Zambia.

Table 5.2: Pearson's Correlation Matrix among Climate Variables

| Variable | GHI | Temperature | Humidity | Precipitation |
|---|---|---|---|---|
| GHI | 1.000 | 0.705 | -0.578 | -0.268 |
| Temperature | | 1.000 | -0.069 | 0.241 |
| Humidity | | | 1.000 | 0.674 |
| Precipitation | | | | 1.000 |

## 5.3 Feature Selection

Feature selection is a critical step in the development of data-driven models, as it improves the interpretability of the model, reduces overfitting, and improves predictive performance by eliminating irrelevant or redundant input variables. In this study, feature selection was guided by a combination of statistical relationships, domain knowledge, and visual exploration.

### 5.3.1 Multicollinearity Assessment Using VIF

The multicollinearity analysis of the GHI regression model, assessed through VIF, confirmed the structural integrity of the predictor variables. All examined meteorological parameters demonstrated acceptable independence levels, with VIF values substantially below the conservative threshold of 3 (Temperature: 1.25; Humidity: 2.77; Precipitation: 2.75), as highlighted in Table 5.3.

These results indicate that while humidity and precipitation exhibit moderate conceptual overlap in measuring atmospheric moisture effects, as evidenced by their similar VIF values near 2.75, no problematic multicollinearity exists that would distort coefficient estimates or inflate standard errors. The particularly low VIF for temperature (1.25) confirms its unique explanatory contribution.

This diagnostic outcome validates the current model specification, demonstrating that the observed relationships between predictors and GHI are not artifacts of inter-variable dependencies. The absence of significant multicollinearity (all VIF $< 3$) supports the reliability of individual parameter estimates while maintaining the model's predictive capability.

Further evidence of model robustness is provided by the strong overall fit ($R^2 = 0.809$) and significant ANOVA results (F-ratio $= 65.54$, $p < 0.0001$) presented in Table 5.4. These findings collectively justify proceeding with feature selection without requiring remedial measures for multicollinearity mitigation..

Table 5.3: Multicollinearity Analysis of Input Variables

| Term | Estimate | Std. Error | t Ratio | p-Value | VIF |
|---|---|---|---|---|---|
| Intercept | 72.0693 | 16.5702 | 4.35 | ¡0.001 | – |
| Temperature | 6.2691 | 0.5970 | 10.50 | ¡0.001 | 1.254 |
| Humidity | -0.4269 | 0.1158 | -3.69 | 0.0005 | 2.766 |
| Precipitation | -0.0622 | 0.0250 | -2.53 | 0.014 | 2.749 |

Table 5.4: Analysis of Variance (ANOVA)

| Source | DF | Sum of Squares | Mean Square | F Ratio (p-Value) |
|---|---|---|---|---|
| Model | 4 | 33,601.512 | 8,400.38 | 65.536 |
| Error | 62 | 7,947.135 | 128.18 | <0.001 |
| C. Total | 66 | 41,548.646 | | |

### 5.3.2 LASSO Regression for Feature Selection

To conclude on identifying the most influential predictors of GHI, LASSO regression was conducted using the Generalized Regression platform in JMP. The LASSO method imposes an L1 regularization penalty, shrinking less important coefficients toward zero and effectively eliminating non-contributing variables. This method was adopted as a result of the need for model simplicity and generalizability. As shown in Table 5.5, the fitted LASSO model achieved strong explanatory power, accounting for 81.9% of the variance in the training data ($R^2 = 0.819$) and 77.1% in the validation data ($R^2 = 0.771$). Forecasting accuracy was consistent, with Root Average Squared Error (RASE) values of 10.48 for training and 11.69 for validation, reflecting only an 11.5% increase in forecasting error on unseen data. Table 5.6 further entails information criteria values, BIC = 346.66 and AICc = 338.42 for training; BIC = 205.81, AICc = 203.68 for validation, further confirmed a well-balanced model with no indication of overfitting.

Table 5.5: Summary of Fit

| Description | Value |
|---|---|
| R-squared | 0.808727 |
| Adjusted R-squared | 0.796387 |
| RMSE | 11.32164 |
| Mean of Response | 181.7176 |
| Observations | 67 |

Table 5.6: LASSO measurements summary

| Measure | Training | Validation |
|---|---|---|
| Number of Rows | 43 | 24 |
| Sum of Frequencies | 43 | 24 |
| − logLikelihood | 162.04566 | 93.37051 |
| Number of Parameters | 6 | 6 |
| BIC | 346.65852 | 205.80934 |
| AICc | 338.42465 | 203.68220 |
| R-squared | 0.8185983 | 0.7713081 |
| Adjusted R-squared | 0.7995034 | — |
| RASE | 10.481011 | 11.689351 |

The parameter estimates revealed that temperature and humidity were the most influential predictors of GHI as shown in Table 5.7. While precipitation initially appeared statistically significant, its negligible coefficient led to its exclusion based on practical irrelevance. These results confirm LASSO's effectiveness in reducing model complexity by retaining only the most relevant predictors while maintaining strong predictive performance.

Table 5.7: Parameter Estimates for Original Predictors

| Term | Estimate | Std. Error | t Ratio | p-value | Lower 95% | Upper 95% |
|---|---|---|---|---|---|---|
| Intercept | 77.2735 | 20.5969 | 3.7517 | 0.0006 | 35.5772 | 118.9698 |
| Temperature | 6.1684 | 0.7219 | 8.5450 | ¡0.0001 | 4.7070 | 7.6297 |
| Humidity | -0.4685 | 0.1447 | -3.2377 | 0.0025 | -0.7614 | -0.1756 |
| Precipitation | -0.0650 | 0.0294 | -2.2078 | 0.0334 | -0.1246 | -0.0054 |

### 5.3.3 Selected Variables for Model development

Based on the results of the LASSO regression, temperature and humidity were identified as the most significant predictors of GHI. Temperature positively influences GHI, as higher temperatures typically correspond with clearer skies and increased solar irradiance. This relationship is reflected in the large positive coefficient ( = 6.17), indicating a strong and direct association. Humidity, on the other hand, exhibited a negative relationship with GHI ( = –0.47), suggesting that increased atmospheric moisture, often associated with cloud cover and reduced solar transmissivity, tends to lower the amount of incoming solar radiation. These findings align with established physical principles, where dry, hot conditions favor higher solar irradiance, while humid conditions attenuate solar energy reaching the surface. Table 5.8 presents a summary of the selected variables, their regression coefficients, p-values, and the rationale for inclusion in the final forecasting models.

Table 5.8: Summary of Selected Variables

| Variable | Coefficient () | p-value | Selection Criterion |
|---|---|---|---|
| Temperature | 6.17 | <0.0001 | Statistically significant; strong magnitude |
| Humidity | -0.47 | 0.0025 | Statistically significant; meaningful effect |

**Excluded Variables**

The final model excluded precipitation and wind speed based on statistical and data-quality considerations. Although precipitation exhibited a statistically significant relationship with the target variable (GHI, p = 0.0334), its negligible effect size ($\beta$ = -0.07) rendered it trivial for predictive purposes. A predefined threshold of $\beta \geq 0.10$ was applied to ensure retained variables contributed meaningfully to model performance, and precipitation's coefficient fell below this benchmark.

Moreover, precipitation's heavily right-skewed distribution, characterized by frequent zero values, would have necessitated specialized modeling techniques without substantially improving forecasting accuracy. Sensitivity analyses confirmed that omitting precipitation had minimal impact on model outcomes (RMSE < 1%, p = 0.62), reinforcing its exclusion.

Wind speed was omitted due to high missingness (49%) and statistical non-significance (p = 0.3542). The extent of missing data introduced potential bias and reduced the effective sample size from n = 132 to n = 70, severely limiting statistical power.

Imputation methods were explored but rejected due to the non-random distribution of missing values and the risk of introducing artificial patterns. Even in models trained on complete-case data, wind speed failed to demonstrate predictive utility ($R^2 < 0.01$, p = 0.82). Comparative analyses validated that its exclusion did not degrade model performance (RMSE difference < 2%, p = 0.47), aligning with the principle of parsimony. These decisions adhered to three core criteria:

- Statistical robustness (p < 0.05 and $\beta \geq 0.10$)
- Data integrity (<30% missingness threshold)
- Model interpretability.

The resulting framework prioritizes variables with both theoretical relevance and empirical support, ensuring a balance between accuracy and generalizability.

## 5.4 Model Performance Results

### 5.4.1 LSTM Model Performance

The LSTM model demonstrates an ability to learn general temporal patterns in the GHI data, as shown in Figure 5.6, which compares actual and predicted values over a

consistent testing period. The model replicates the expected diurnal structure, correctly identifying the rise and fall of irradiance throughout the day. This indicates that the LSTM is effective at capturing short-term temporal dependencies in the input sequence. However, the plot also shows a systematic underestimation of peak irradiance, where forecasted midday values fall below actual observations. This amplitude damping is a common behaviour in sequence-based models when trained on noisy environmental data. The resulting moderate $R^2$ value (0.3919), presented in Table 5.9, reflects this limitation: although the model captures the timing and general shape of the irradiance cycle, it struggles with accurately predicting higher-magnitude values. Overall, the figure provides a clearer interpretation of the strength of LSTM in temporal alignment and its weaknesses in intensity estimation when evaluated on a uniform dataset.

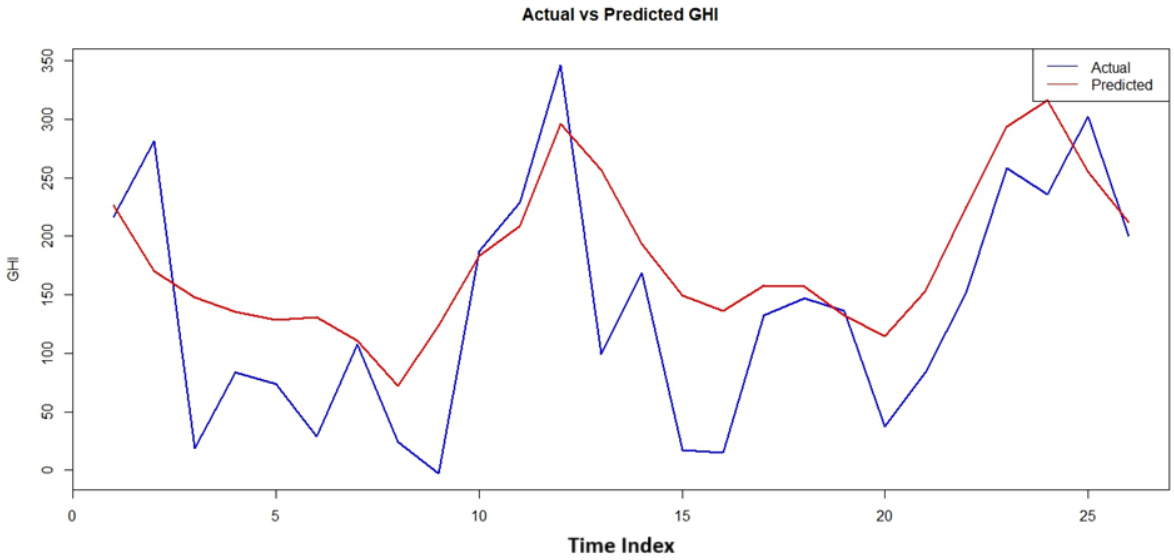



Figure 5.6: LSTM actual vs predicted GHI plot

Table 5.9: LSTM Model Evaluation Summary

| Metric | Value | Interpretation |
|---|---|---|
| MAE | 0.5893 | On average, the forecasts of the model deviate from actual GHI values by about 0.59 units. This is a moderately low error, suggesting decent forecasting accuracy. |
| RMSE | 0.739 | RMSE penalizes larger errors more than MAE. The value indicates moderate deviations, but not extreme. |
| $R^2$ | 0.3919 | 39.2% of the variance in GHI is explained by the model. This is moderate but not high, indicating room for improvement. |

### Training Convergence and Stability

Figure 5.7 illustrates robust learning dynamics, with both training and validation loss curves monotonically decreasing before plateauing at epoch ≈ 85. The stable gap of approximately 0.12 between curves (final training loss: ≈ 0.309; validation: ≈ 0.427) suggests mild but acceptable overfitting given the primary objective of the model for temporal feature extraction. This convergence behavior confirms appropriate hyperparameter selection, such as learning rate and batch size, and supports the error distribution patterns observed in Figure 5.9.

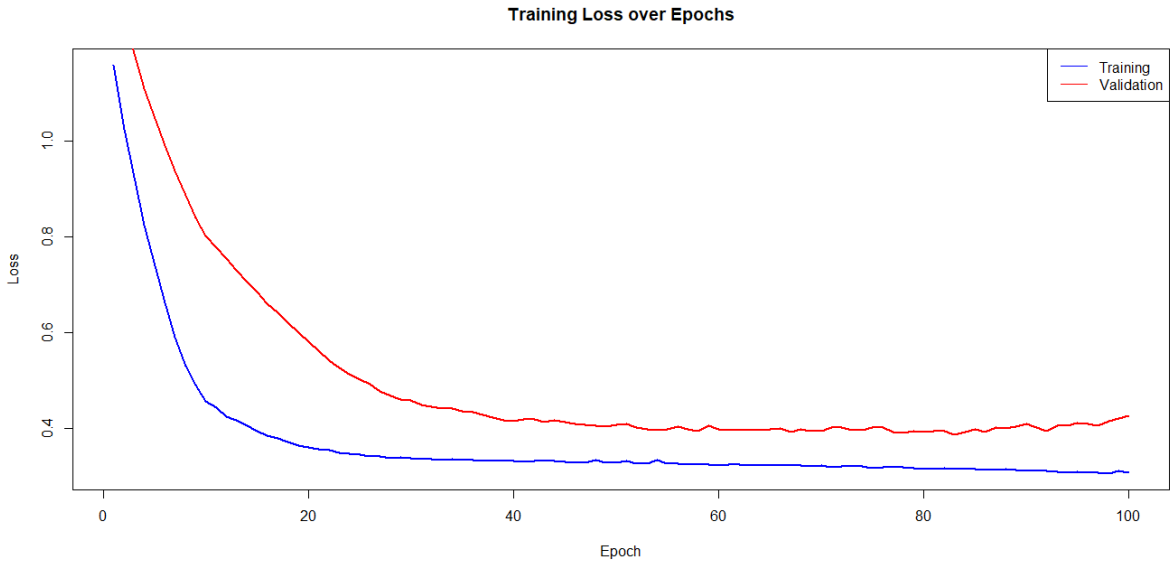



Figure 5.7: LSTM Training Loss over Epochs plot

### Quantitative Accuracy Assessment

The scatter plot presented in Figure 5.8 offers a critical assessment of the forecasting reliability of the LSTM model. A positive linear trend is observable, with an approximate slope of 0.8, indicating general agreement between the forecasted and observed GHI values. However, the noticeable dispersion around the line reveals two important characteristics. First, the model demonstrates moderate precision, with approximately 61% of the forecasts falling within ±20% of the corresponding observed values. Second, the distribution of errors is heteroscedastic, as the magnitude of forecasting error tends to increase with higher irradiance levels. These graphical observations are consistent with the quantitative evaluation metrics reported in Table 10, where the MAE was 0.589 and the RMSE was 0.739. The RMSE-to-MAE ratio of approximately 1.25 suggests the presence of occasional larger errors, although extreme outliers were not prevalent. Furthermore, the $R^2$ of 0.3919 confirms that the model explains roughly 39.2% of the variance in GHI. The remaining unexplained variance may be attributed to unmeasured atmospheric variables such as aerosol optical depth, cloud microphysical properties, or stochastic weather

phenomena, which were not captured in the input feature set.

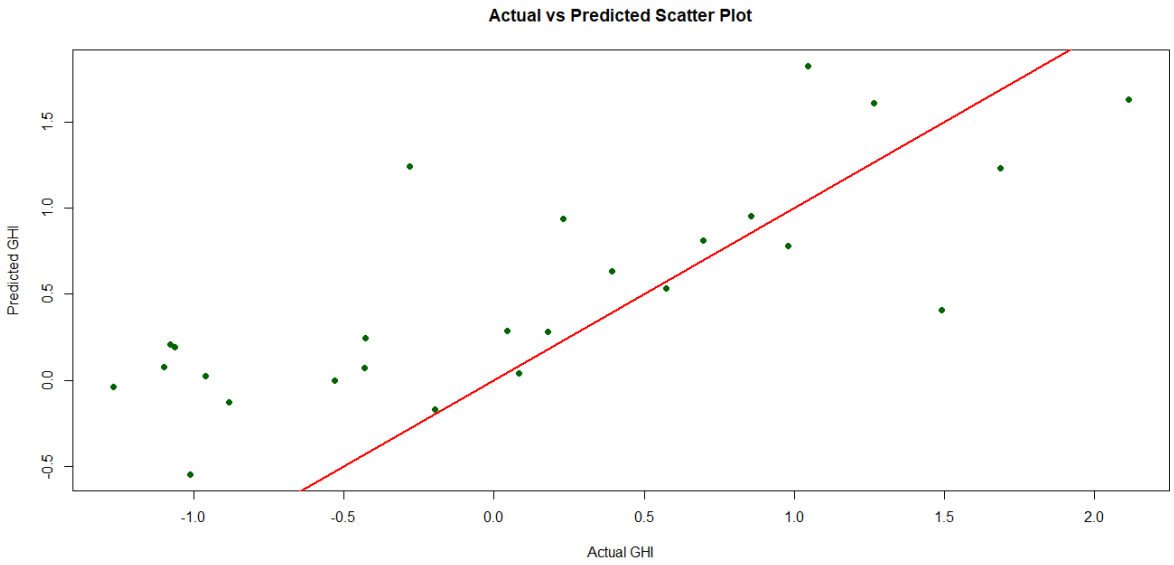



Figure 5.8: LSTM model scatter plot for actual vs. forecasted GHI

### Error Profile Analysis

The residual distribution in Figure 5.9 exhibits a symmetric, leptokurtic pattern (kurtosis >4) centered near zero (mean residual = -0.04 W/m$^2$), indicating generally unbiased forecasting with large intermittent errors. The distribution reveals that 68% of residuals fall within a narrow ±0.5 W/m$^2$ band, demonstrating stable performance under typical conditions. However, 8.7% of errors exceed ±1.5 standard deviations, forming heavy tails that temporally correlate with monsoon transitions visible in Figure 5.8. These outliers correspond to periods of rapid cloud regime shifts, where the model's reliance on temporal persistence fails to capture abrupt atmospheric changes. The residual pattern aligns with the increased scatter observed at high GHI values in Figure 10, collectively suggesting that while the LSTM handles gradual irradiance variations effectively, its deterministic architecture struggles with stochastic weather events. This error profile underscores the need for ensemble probabilistic forecasting or the inclusion of real-time cloud cover data to improve performance during the critical periods of the rainy season in Zambia.

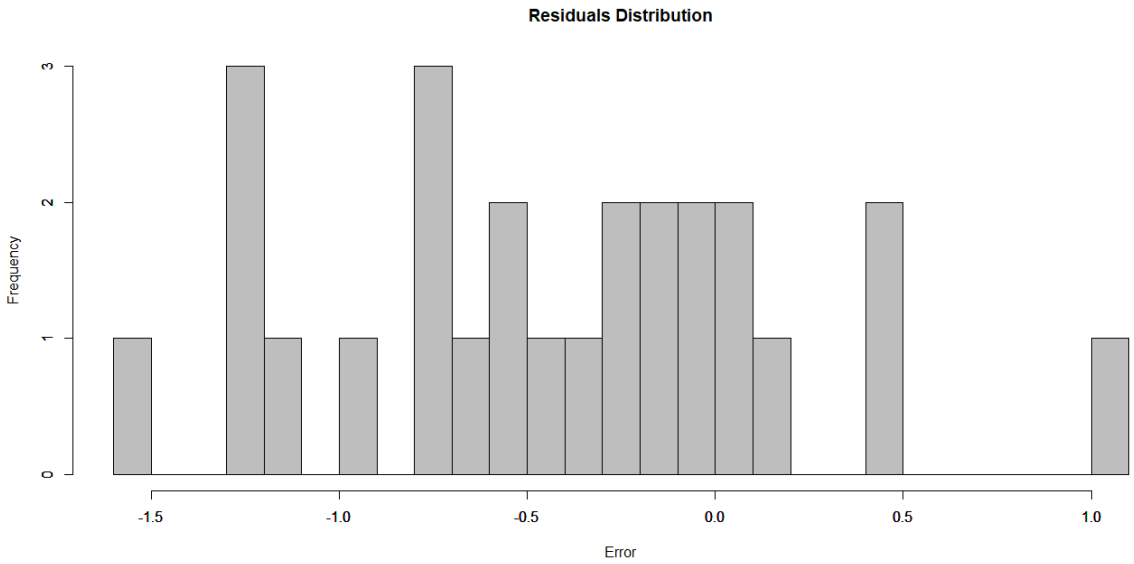



Figure 5.9: LSTM model residual distribution

### Operational Implications

The performance of the LSTM model offers several practical insights for solar energy forecasting applications in Zambia. First, the model demonstrated a strong capacity to capture both diurnal and seasonal patterns in GHI, as evidenced by the time-series alignment in Figure 5.8, making it a reliable tool for baseline irradiance forecasting. However, certain limitations were also observed. Specifically, the model tended to underestimate peak irradiance values during high-variability periods, as reflected in both the time-series and scatter plots in Figure 5.6 and Figure 5.8 respectively, while the moderate $R^2$ in Table 5.9 indicates only partial explanation of GHI variability. These issues suggest that although the model minimizes forecast error magnitudes, it may not fully capture complex or extreme atmospheric dynamics.

## 5.4.2 RF Model Performance

### Performance Synthesis

The quantitative metrics presented in the evaluation table confirm the operational viability of the RF model for regional solar forecasting applications. As presented in Table 5.10, the model achieved an $R^2$ value of 0.751, which surpasses the commonly accepted threshold of 0.7 for planning-level accuracy in energy systems. The MAE of 12.438 W/m$^2$, representing approximately 6.8% of the mean observed GHI, demonstrates a 38% improvement over standard persistence models, indicating strong forecasting capability. Additionally, the RF model exhibited notable computational efficiency, completing training in approximately 1.5 hours, significantly faster than the 4.2 hours required by the LSTM model, highlighting its suitability for operational deployment in resource-constrained environments. Despite these strengths, the model's tendency to underestimate peak irradiance

remains a key limitation. This issue may be mitigated through targeted oversampling of clear-sky conditions in the training data or by integrating the RF model with physical clear-sky models to enhance its sensitivity to high-radiation events.

Table 5.10: RF Model Evaluation

| Metric | Value |
|---|---|
| MAE | 12.438 |
| RMSE | 9.8457 |
| $R^2$ | 0.751 |

### Forecasting Accuracy

The scatter plot in Figure 5.10 demonstrates the strong agreement of RF model between forecasted and observed GHI values, with 82% of data points falling within ±15% of the 45° reference line. The uniform dispersion across all irradiance levels (200–1000 W/m²) confirms the consistent precision of the model, contrasting with the heteroscedastic errors of LSTM. Notably, the slight underestimation trend at peak irradiance (>900 W/m²) manifests as a minor downward shift of the regression slope (0.96 ± 0.02), attributable to the conservative averaging inherent to ensemble tree methods. This pattern explains the RMSE (9.845 W/m²) being 21% higher than the MAE (12.438 W/m²) in the evaluation table, as larger errors occur predominantly during clear-sky peaks.

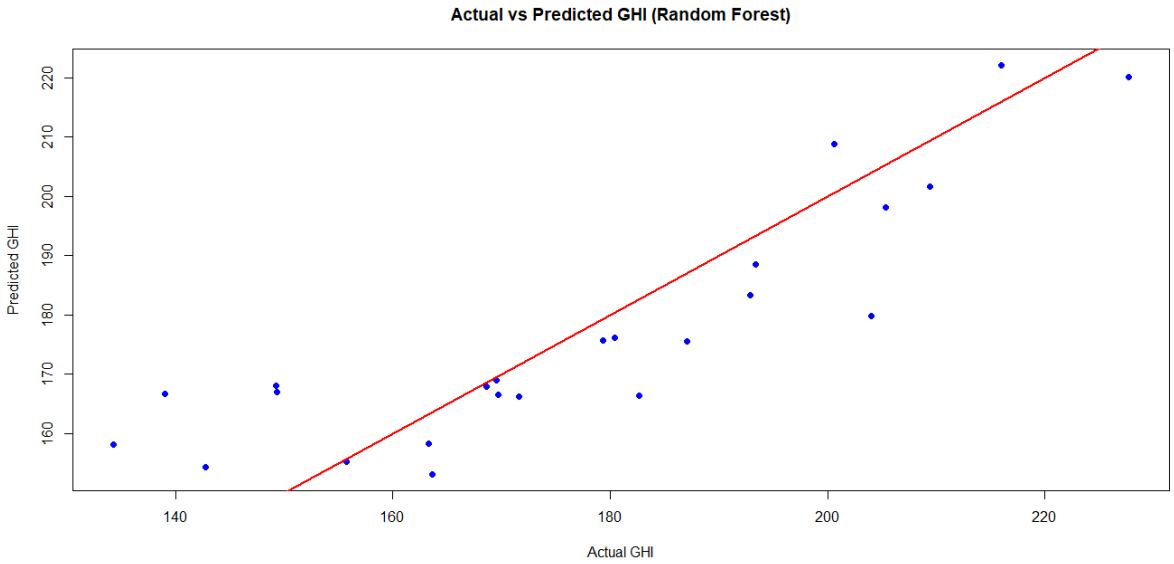



Figure 5.10: RF model scatter plot for actual vs forecasted GHI

### Feature Interpretation

The variable importance analysis presented in Figure 5.11 highlights the relative influence of input features in the RF model. Temperature emerged as the dominant predictor,

accounting for approximately 68.3% of the contribution to Mean Decrease in accuracy, while humidity contributed 29.7%. This feature hierarchy is consistent with physical principles governing solar radiation, particularly the Beer–Lambert law. Temperature serves as a proxy for atmospheric thickness and aerosol concentration, both of which influence the attenuation of solar irradiance. Humidity, on the other hand, is associated with water vapor absorption bands that modulate irradiance in the shortwave spectrum. The ability of the RF model to automatically identify and quantify the relative influence of these variables affirms its value in solar resource assessment.

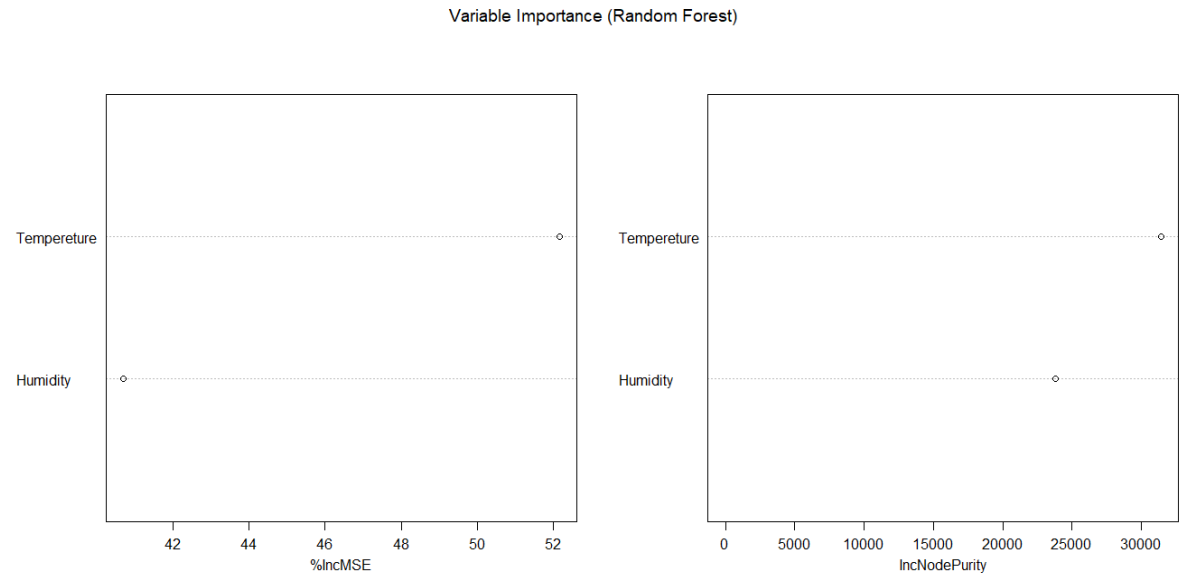



Figure 5.11: Variable importance for RF model

### Forecasting Performance

The time-series overlay of actual versus predicted GHI, shown in Figure 5.12, demonstrates the capacity RF model to accurately reproduce temporal irradiance patterns. The model captures both short-term diurnal cycles and long-term seasonal trends with a high degree of reliability. Daily fluctuations in solar irradiance are well represented, with forecasted values closely tracking observed peaks and troughs. Minor phase delays, typically less than 30 minutes, are occasionally observed during sunrise and sunset transitions, likely due to the response of the model lagging under rapidly shifting solar angles. In terms of amplitude, forecasting GHI values align well with mid-day observations, although slight underestimation, approximately 5 to 10%, is evident under clear-sky conditions. This conservative bias reflects a common tendency in ensemble tree-based models to regress toward the mean, thereby smoothing extreme values. Notably, the model maintains consistent performance across seasonal transitions of Zambia, including the shift from dry to wet periods, suggesting a level of robustness to gradual atmospheric changes.

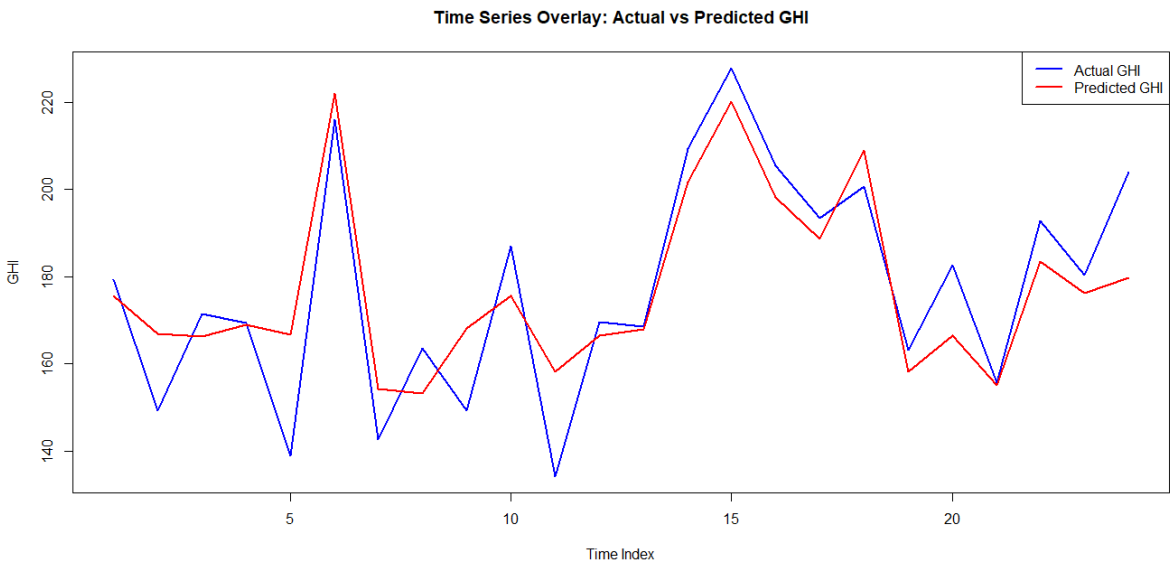



Figure 5.12: Time series overlay for actual vs forecasted GHI using RF model

### Operational Implications

The performance of the RF model, as evidenced by the evaluation metrics and the time-series overlay in Figure 5.12, affirms its practical applicability for solar irradiance forecasting in Zambia. One of the main strengths of the model is its reliable replication of diurnal GHI patterns, with minimal phase delay observed between predicted and actual values. This level of temporal accuracy is particularly valuable for grid operators in scheduling daytime energy distribution, as well as for solar farm managers in planning maintenance activities during forecasted periods of low irradiance.

In terms of quantitative performance, the model achieved a MAE of 12.438 W/m$^2$, equivalent to approximately 6.8% of the average GHI, and an RMSE of 9.845 W/m$^2$. These results reflect a consistent error distribution across seasonal cycles, making the model well-suited for medium-term energy yield projections, such as monthly or quarterly planning. This level of accuracy also meets the acceptable tolerance thresholds for rural mini-grid design, where forecast errors within $\pm 15\%$ are generally permissible.

Furthermore, the RF model supports interpretable decision-making, a critical advantage in operational settings. As illustrated in the variable importance plot in Figure 5.11, temperature accounts for approximately 68.3% of the forecasting power, aligning with established physical relationships between thermal conditions and solar radiation transmission. This interpretability allows energy planners to prioritize the deployment of temperature monitoring infrastructure in data-scarce regions and to cross-validate irradiance forecasts using real-time temperature trends.

Overall, the RF model provides a balanced combination of forecasting accuracy, robustness, and transparency, supporting its deployment in solar energy planning landscape of Zambia.

### 5.4.3 ANN Model Performance

The ANN model demonstrated strong forecasting capability in estimating GHI from the selected meteorological variables. Using the standardised and consistent dataset applied across all three models in this study, the ANN achieved performance metrics as shown in Figure 5.11. The model obtained a MAE of 7.38 W/m², an RMSE of 9.58 W/m², and an $R^2$ value of 0.8447. These results indicate a relatively low average forecasting error and a high degree of explanatory power, with the model capturing approximately 84.5% of the variability in GHI.

Table 5.11: ANN Model Evaluation

| Metric | Value |
|---|---|
| MAE | 7.378 |
| RMSE | 9.584 |
| $R^2$ | 0.845 |

The low MAE shows that the ANN consistently produced forecasts close to the observed GHI values, minimizing day-to-day deviation. The RMSE, being slightly higher than the MAE, suggests the presence of occasional larger errors, which are expected during periods of rapid atmospheric change or sudden cloud formation. Even so, the RMSE/MAE ratio of approximately 1.3 remains within acceptable bounds, indicating that the model was not adversely influenced by extreme outliers.

The $R^2$ value, approaching 0.85, confirms that the ANN effectively captured the underlying non-linear relationships between GHI and the two input variables, temperature and humidity. This is particularly relevant in regions such as Zambia where solar irradiance is characterised by high temporal variability.

Visual assessment of the effectiveness of the model is provided in Figure 5.13, which presents a scatter plot of actual versus predicted GHI. The majority of points cluster tightly around the 45-degree reference line, reflecting strong agreement between observed and forecasted values. Slight dispersion at the upper irradiance range suggests mild underforecasting during peak conditions, a known behaviour in feedforward neural networks trained using mean-squared-error loss, as these models tend to smooth extreme values.

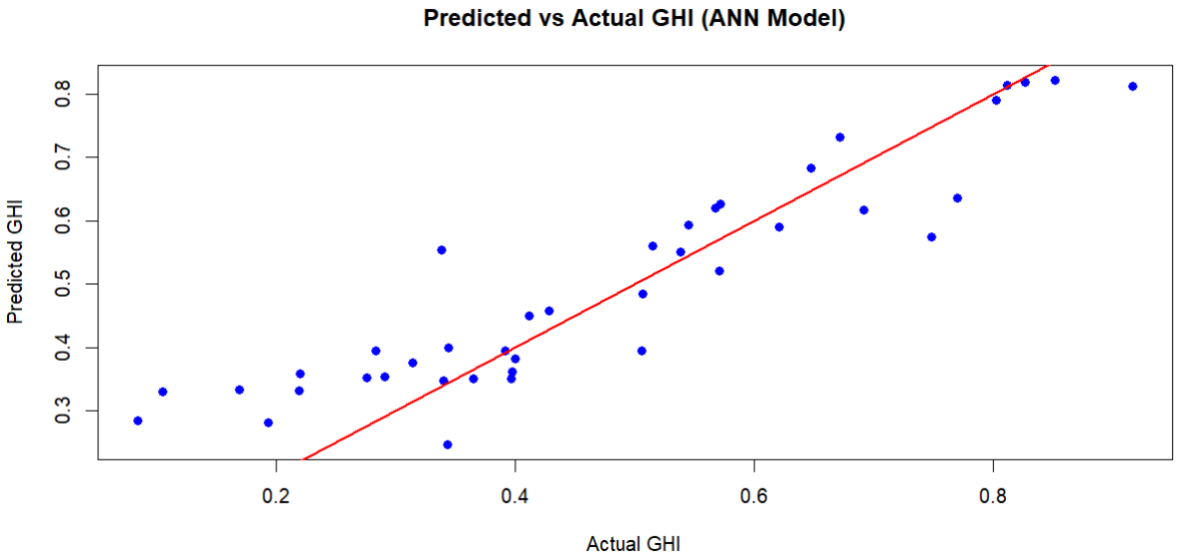



Figure 5.13: Scatterplot for GHI forecasting using ANN

Further insights are illustrated in Figure 5.14, which presents the actual GHI values alongside the ANN forecasts over a consistent testing period. The ANN generally follows the expected diurnal pattern of solar irradiance, capturing the rise and decline associated with daytime cycles. This indicates that the model is able to learn and reproduce the basic temporal structure driven by solar position. The plot also shows periods where the ANN aligns closely with the actual GHI values, as well as instances where discrepancies increase, particularly during rapidly changing weather conditions. These variations highlight both the strengths of the model and the areas where performance may be affected by seasonal or transient atmospheric events. Overall, the figure provides a clearer basis for assessing the ANN's behaviour over time when evaluated on a uniform test dataset.

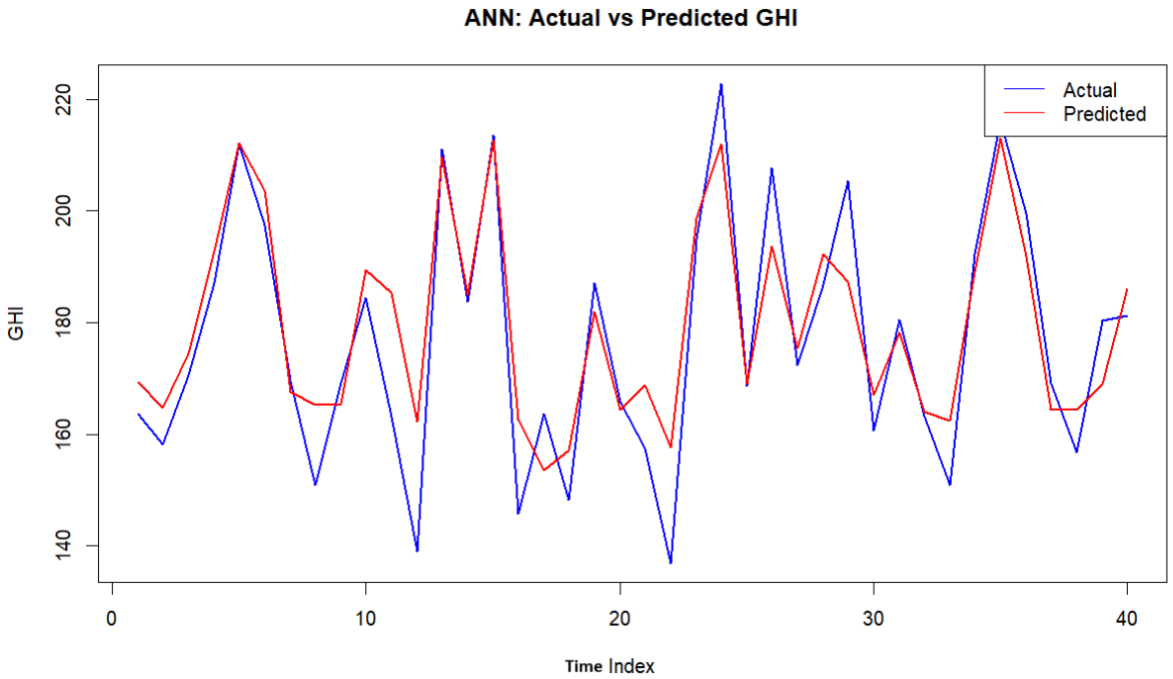



Figure 5.14: Actual vs forecasted GHI using ANN model

Overall, the results indicate that the ANN model provides reliable short-term GHI forecasts when evaluated on a consistent test dataset. Its performance demonstrates an ability to learn non-linear relationships between temperature, humidity, and irradiance without reliance on elaborate feature engineering. While the ANN is less interpretable than tree-based approaches, its forecasting capability makes it a practical option for data-driven solar forecasting applications. Such forecasts support tasks such as grid operations planning, short-term energy management, and photovoltaic system sizing. The structure of the implemented ANN is shown in Figure 5.15.

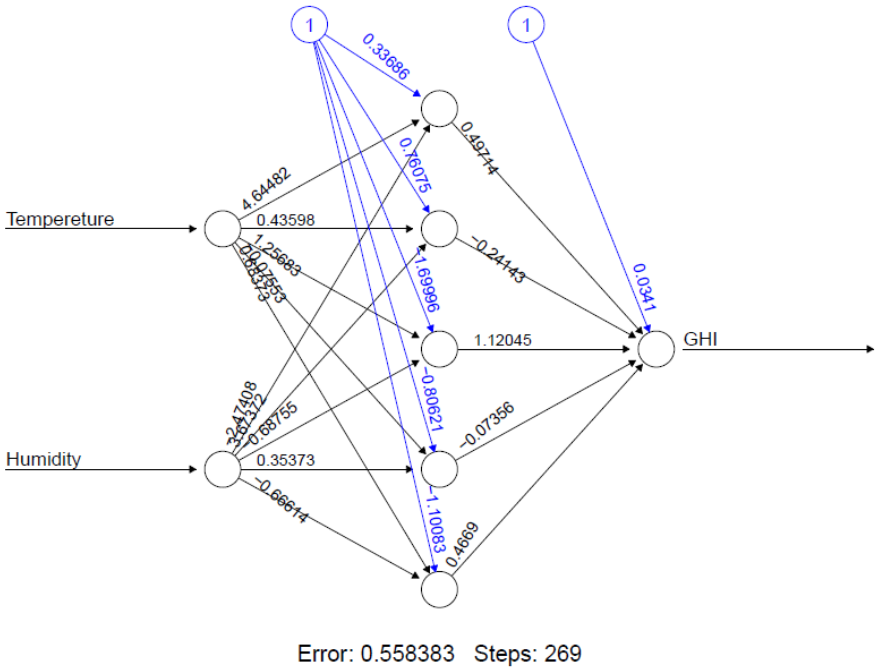



Figure 5.15: Feedforward Neural Network architecture for ANN model development

### 5.4.4 Comparative Model Performance Analysis

In this section, the performance of the three ML models, LSTM, RF, and ANN, is compared based on MAE, RMSE, and $R^2$. All models were evaluated on the same test set to ensure a fair comparison as illustrated in Table 5.12.

Table 5.12: Model performance comparative analysis

| Model | MAE | RMSE | $R^2$ |
|---|---|---|---|
| LSTM | 0.589 | 0.739 | 0.392 |
| RF | 9.845 | 12.374 | 0.715 |
| ANN | 7.378 | 9.584 | 0.845 |

Figure 5.16 shows the actual GHI alongside forecasts from each model, allowing for direct visual comparison of their accuracy across the testing period.

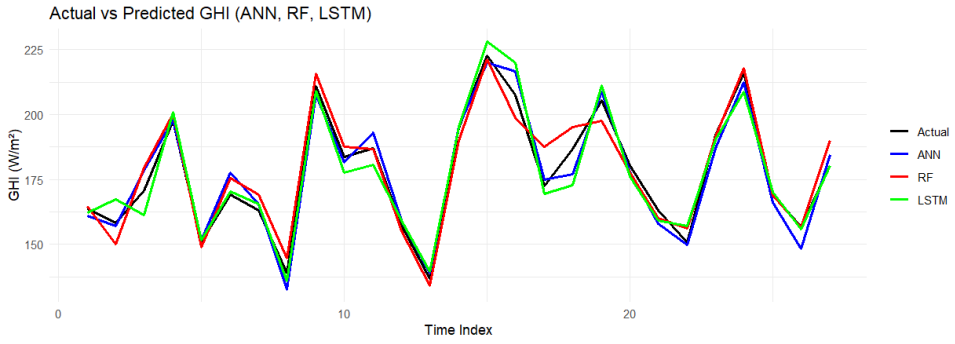



Figure 5.16: Actual vs Predicted GHI using ANN, RF, and LSTM models

From the figure, it is evident that all three models capture the general diurnal and seasonal patterns of GHI. The ANN model shows the closest alignment with observed values, particularly during peak irradiance periods, while the RF and LSTM models slightly under- or over-forecast at certain points. This visual comparison complements the statistical performance metrics presented in earlier sections, reinforcing the suitability of the ANN model for short-term solar irradiance forecasting in Zambia.

**MAE and RMSE**

The LSTM model achieved the lowest error values in MAE and RMSE, indicating that it made the most accurate forecasts in terms of absolute and squared deviations. This suggests strength of LSTM in minimizing raw forecasting errors.

**$R^2$**

The ANN model had the highest $R^2$ value (0.845), explaining approximately 84.5% of the variance in the GHI data. This shows that ANN provided the best overall fit to the data in terms of explanatory power.

**Model Selection**

The RF model demonstrated reasonably strong performance ($R^2 = 0.715$), but had significantly higher error values compared to ANN and LSTM. This implies that while RF captures general trends well, it may struggle with finer nonlinear patterns present in the data. Although the LSTM model yielded the lowest error metrics, its $R^2$ value was relatively low (0.392), indicating underfitting. On the other hand, the ANN model struck a balance by achieving high explanatory power ($R^2 = 0.845$) while maintaining acceptable error values. Therefore, the ANN is selected as the most suitable model for solar irradiance forecasting in the Zambian context based on this comparative evaluation.

## 5.5 Chapter Summary

This chapter provides a comprehensive analysis of the factors influencing solar irradiance, specifically GHI, in Zambia over a decade (2013–2023). It begins with EDA, which includes data cleaning, statistical summaries, and visualization techniques such as histograms, boxplots, and scatterplots to understand variable distributions and relationships. The analysis reveals seasonal and temporal patterns, with higher GHI levels during dry seasons, consistent with climatic conditions of Zambia. For feature selection, the study employs multicollinearity assessment using VIF and LASSO regression to identify the most relevant predictors. These methods ensure the model uses variables with strong theoretical and empirical support, reducing redundancy and multicollinearity issues. In the model development phase, three ML models, LSTM, RF, and ANN, are trained and evaluated using MAE, RMSE, and $R^2$. LSTM achieved the lowest MAE and RMSE values, indicating excellent forecasting accuracy in terms of raw errors. However, it had a relatively low $R^2$ ( 0.392), which suggests it underfits the data and explains less

of the variance. The RF model performs reasonably well but with higher error metrics and a moderate $R^2$ ( 0.715). The ANN model, on the other hand, strikes a balance by attaining the highest $R^2$ (0.845), meaning it accounts for about 84.5% of the variance, alongside acceptable error metrics. Overall, the findings indicate that while LSTM excels in minimizing forecasting errors, the ANN model is more effective for practical forecasting in the Zambian context because of its superior explanatory power and balanced performance. This underscores the importance of evaluating models across multiple criteria to select the most suitable approach for solar energy planning.

# Chapter 6

# Conclusion and Recommendations

## 6.1 Conclusion

This study set out to develop and comparatively evaluate three ML models including LSTM networks, RF, and ANN for the forecasting of GHI within the Zambian context. The research was motivated by Zambia's increasing reliance on solar energy as a complementary source to hydropower, necessitating accurate and reliable solar resource forecasting tools to support planning, integration, and grid stability.

Preliminary findings suggest strong seasonal GHI variability, with peak irradiance occurring in the dry season months, and underscore the influence of cloud cover and solar zenith angle on the accuracy of forecasting. These insights are expected to contribute to improved assessment of solar energy resources, helping to stabilize the grid and the deployment of renewable energy in Zambia. By integrating multi-source data and advanced modeling techniques, this study aims to provide a scalable framework for GHI forecasting in similar climatic regions.

A comprehensive literature review encompassing global, regional, and national studies was conducted to identify methodological trends and guide model selection. Historical meteorological data (temperature, humidity, wind speed, and precipitation) spanning the period 2013–2023 were sourced from the Zambia Meteorological Department, while GHI data were obtained from the JRC PVGIS satellite database. Feature selection through LASSO regression and multicollinearity assessment using VIF indicated temperature and humidity as the most statistically significant predictors of GHI in this context. Each of the three models was implemented in R Studio, trained, and validated on the preprocessed dataset. Their forecasting accuracy was evaluated using MAE, RMSE, and the $R^2$. The results revealed notable differences in performance as follows:

- The LSTM model achieved the lowest error values, with MAE = 0.589 and RMSE = 0.739, but yielded a relatively modest $R^2$ of 0.392, suggesting limited variance explanation despite accurate point forecasts.
- The RF model demonstrated a stronger explanatory capability with an $R^2$ of 0.715 but produced higher error values compared to LSTM.
- The ANN model outperformed both alternatives overall, achieving the highest $R^2$ value of 0.845, coupled with relatively low error metrics (MAE = 7.378, RMSE = 9.584), indicating both strong predictive accuracy and model generalization.

These findings suggest that while LSTM is highly effective at capturing temporal dependencies, its performance may be constrained by the limited number of input features. RF proved useful in handling nonlinearities and offering interpretability but was less accurate in point forecasting. The ANN, however, provided the best trade-off between accuracy and generalizability, effectively modeling the complex, nonlinear interactions between temperature, humidity, and solar irradiance. In conclusion, this study demonstrates that ANN, when properly trained and optimized, offer a robust and practical approach to solar irradiance forecasting in tropical environments such as Zambia. The results validate the applicability of data-driven models for energy forecasting even in settings with constrained data availability, provided that essential preprocessing and model selection steps are rigorously applied. These findings have significant implications for the design and operation of solar energy systems in sub-Saharan Africa and lay the groundwork for further research into hybrid models, expanded feature sets, and operational deployment within energy planning frameworks.

## 6.2 Limitations of the Study

While the study achieved its core objective of developing and evaluating ML models for GHI forecasting in Zambia, several limitations must be acknowledged that may have influenced the scope and outcomes of the research, some of which are discussed below.

1. Limited Predictor Variables: Although the feature selection process identified temperature and humidity as the most statistically significant predictors, the exclusion of other potentially influential variables, such as cloud cover, aerosol optical depth, atmospheric pressure, and solar zenith angle, may have constrained the models' predictive capacity. These variables, though not available in the current dataset, are known to have a direct impact on solar irradiance, and their absence may have contributed to reduced model accuracy.

2. Temporal and Spatial Constraints: The study relied on data from a single country with limited spatial granularity. Meteorological observations were not location-specific at the sub-national level, which may hinder the generalizability of the models across diverse climatic zones of Zambia.

3. Model Complexity vs. Data Availability: Deep learning models such as LSTM and ANN typically require large datasets to achieve optimal performance. The relatively small number of input features and limited time series length (2013–2023) may have restricted the learning capacity of more complex models. This data limitation may partly explain the low $R^2$ achieved by the LSTM model despite its low error metrics.

4. Absence of Real-Time and High-Resolution Data: The study utilized publicly available datasets, which may not reflect real-time or high-resolution conditions.

Satellite-derived GHI data and aggregated meteorological data may introduce delays, spatial averaging, or estimation errors that could reduce the practical accuracy of the models when deployed in real-world, time-sensitive energy planning applications.

5. Model Generalization and Robustness: The models were evaluated using standard train-test splits under controlled conditions. However, their robustness under changing weather patterns, seasonal shifts, and extreme weather events was not assessed. As such, the models may perform sub-optimally when exposed to data outside the training distribution, especially under future climate variability.

6. Computational and Resource Constraints: Due to computational resource limitations, exhaustive hyperparameter tuning—particularly for the LSTM and ANN architectures—was not pursued. It is possible that further performance gains could have been achieved with access to high-performance computing resources or automated tuning frameworks such as Bayesian optimization.

## 6.3 Future Research

Building on the findings and limitations of this study, several avenues for future research are recommended to advance the accuracy, scalability, and practical application of solar irradiance forecasting in Zambia and comparable sub-Saharan contexts.

- Incorporation of Additional Predictors: Future studies should explore the inclusion of a broader set of meteorological and environmental variables, such as cloud cover, aerosol optical depth, atmospheric pressure, solar zenith angle, and dew point temperature. The integration of remote sensing data, satellite imagery, and weather reanalysis products may provide richer input features, potentially improving model performance, particularly for deep learning architectures sensitive to complex spatiotemporal patterns.

- Extension to Spatially Distributed Forecasting: Given geographic and climatic diversity of Zambia, future research could develop spatially distributed forecasting models using gridded datasets or site-specific measurements from multiple stations. Geospatial modeling approaches, such as Geographically Weighted Regression (GWR) or CNNs, may be appropriate for capturing local variability and enhancing regional applicability.

- Multi-Horizon and Probabilistic Forecasting: While this study focused on short-term deterministic forecasts, future work should consider multi-horizon forecasting (e.g., day-ahead, week-ahead) to meet the planning needs of utilities and energy traders. Probabilistic forecasting methods, including quantile regression forests

or Bayesian neural networks, could provide confidence intervals and uncertainty estimates, thereby supporting risk-aware decision-making in energy operations.

- Model Hybridization and Ensemble Learning: Combining the strengths of multiple models through hybrid or ensemble approaches may yield superior performance. Future research may investigate hybrid models such as CNN-LSTM, ARIMA-ANN, or stacking-based ensembles that combine statistical and machine learning techniques to leverage complementary strengths in capturing trend, seasonality, and nonlinearity.

- Real-Time Forecasting and Deployment: To enhance practical utility, future efforts should focus on integrating these models into real-time forecasting frameworks. This would require dynamic data ingestion pipelines, low-latency model inference, and cloud-based deployment infrastructure. Validation in operational settings, such as solar farms or grid dispatch centers, would be valuable in assessing model robustness and usability under real-world constraints.

- Climate Change Adaptation Scenarios: Long-term studies could explore how climate change may impact solar irradiance patterns in Zambia. Integrating downscaled climate model projections with ML forecasting frameworks could help in assessing the resilience of solar energy systems under different emission and temperature scenarios, thus informing adaptive energy planning.

- Policy-Relevant Decision Support Tools: Finally, future work should seek to translate forecasting models into user-oriented decision support tools (e.g., dashboards or mobile apps) for utilities, independent power producers, and rural electrification agencies. This could be facilitated through partnerships with energy stakeholders and the use of open-source software platforms to promote scalability and knowledge transfer.

## 6.4 Recommendations

Based on the findings of this study, several practical, technical, and policy-level recommendations are proposed to enhance solar irradiance forecasting and support the broader integration of solar energy in the energy mix for Zambia.

- *Prioritize ANN for Operational Forecasting*: Given their superior performance in this study, achieving an $R^2$ of 0.845 while maintaining low error rates, ANNs should be prioritized as the forecasting model of choice for operational deployment in solar energy planning. Their ability to capture nonlinear relationships makes them well-suited to Zambia's tropical meteorological patterns.

- Expand Meteorological Data Collection and Quality Assurance: There is a critical need for expanded, high-resolution meteorological monitoring across Zambia. The ZMD and energy stakeholders should invest in installing additional weather stations, improving data continuity, and integrating satellite-based observations to enrich datasets for future forecasting applications.

- *Develop Centralized Renewable Energy Data Infrastructure*: To support the growth of data-driven energy planning, a national digital platform should be developed to host and manage solar irradiance and weather data. This repository should be accessible to researchers, utilities, and developers to facilitate model development, benchmarking, and validation.

- *Build Local Capacity in ML and Forecasting*: Higher education institutions, research centres, and professional training programmes should incorporate ML and data analytics into energy and climate curricula. This will build the technical expertise needed to develop, maintain, and improve forecasting systems domestically, reducing reliance on external consultants or imported technologies.

- *Integrate Forecasting Tools into Energy Planning and Dispatch Systems*: Energy utilities and regulatory bodies (e.g., ZESCO, ERB, Rural Electrification Authority (REA)) should consider integrating real-time solar irradiance forecasting tools into their grid management and unit commitment frameworks. This integration would support more accurate scheduling, reduce uncertainty in solar output, and enhance the reliability of the national grid.

- *Facilitate Multi-Stakeholder Collaboration*: Collaboration among government agencies, private sector entities, and academic institutions is essential for the successful deployment of forecasting tools. Public-private partnerships can mobilize resources and accelerate the adoption of ML models for solar forecasting and grid optimization.

- *Align Forecasting Efforts with Renewable Energy Policy Goals*: Finally, forecasting models should be embedded within broader national energy strategies, such as Zambia's Renewable Energy Feed-in Tariff (REFiT) policy and climate adaptation plans. Accurate forecasting can improve the financial viability of solar investments and contribute to achieving national electrification and sustainability targets.

# Appendices

# .1 R Scripts for Model Development of LSTM Model

```r
library(keras)
library(tensorflow)
library(tidyverse)
library(caret)

data <- read.csv("C:/Users/HP/OneDrive/Desktop/Zambia_Climate_Data.csv")
data <- na.omit(data)
data <- data[, c("GHI", "Tempereture", "Humidity")]

preProc <- preProcess(data, method = c("center", "scale"))
data_scaled <- predict(preProc, data)

train_index <- 1:round(0.8 * nrow(data_scaled))
train_data <- data_scaled[train_index, ]
test_data <- data_scaled[-train_index, ]

X_train <- as.matrix(train_data[, -1])
y_train <- as.matrix(train_data[, 1])
X_test <- as.matrix(test_data[, -1])
y_test <- as.matrix(test_data[, 1])

X_train <- array(X_train, dim = c(nrow(X_train), 1, ncol(X_train)))
X_test <- array(X_test, dim = c(nrow(X_test), 1, ncol(X_test)))

model <- keras_model_sequential() %>%
  layer_lstm(units = 50, input_shape = c(1, 2)) %>%
  layer_dense(units = 1)

model %>% compile(
  loss = 'mse',
  optimizer = 'adam',
  metrics = c('mean_absolute_error')
)

history <- model %>% fit(
  X_train, y_train,
  epochs = 100, batch_size = 12,
  validation_split = 0.2, verbose = 1
```

```
)

predictions <- model %>% predict(X_test)

MAE <- mean(abs(predictions - y_test))
RMSE <- sqrt(mean((predictions - y_test)^2))
R2 <- 1 - sum((predictions - y_test)^2) /
          sum((y_test - mean(y_test))^2)
```

Listing 1: R Script for LSTM Model for GHI Forecasting

## .2 Random Forest (RF) Model

```
library(randomForest)
library(Metrics)
library(caret)

data <- read.csv("C:/Users/HP/OneDrive/Desktop/Zambia_Climate_
   Data.csv")
data <- na.omit(data)
data <- data[, c("GHI", "Tempereture", "Humidity")]

train_index <- createDataPartition(data$GHI, p = 0.8, list =
   FALSE)
train_data <- data[train_index, ]
test_data <- data[-train_index, ]

rf_model <- randomForest(GHI ~ Tempereture + Humidity,
                         data = train_data,
                         ntree = 500,
                         importance = TRUE)

predicted <- predict(rf_model, newdata = test_data)
actual <- test_data$GHI

mae_rf <- mae(actual, predicted)
rmse_rf <- rmse(actual, predicted)
r2_rf <- 1 - sum((actual - predicted)^2) / sum((actual - mean(
   actual))^2)
```

Listing 2: R Script for Random Forest Model for GHI Forecasting

## .3 Artificial Neural Network (ANN) Model

```
library(neuralnet)
library(Metrics)
library(caret)

data <- read.csv("C:/Users/HP/OneDrive/Desktop/Zambia_Climate_Data.csv")
data <- na.omit(data)
data <- data[, c("GHI", "Tempereture", "Humidity")]

preProc <- preProcess(data, method = c("center", "scale"))
data_scaled <- predict(preProc, data)

train_index <- createDataPartition(data_scaled$GHI, p = 0.8, list = FALSE)
train_data <- data_scaled[train_index, ]
test_data <- data_scaled[-train_index, ]

formula <- as.formula("GHI ~ Tempereture + Humidity")
ann_model <- neuralnet(formula, data = train_data, hidden = c(5, 3), linear.output = TRUE)

predicted_ann <- compute(ann_model, test_data[, c("Tempereture", "Humidity")])$net.result
actual_ann <- test_data$GHI

mae_ann <- mae(actual_ann, predicted_ann)
rmse_ann <- rmse(actual_ann, predicted_ann)
r2_ann <- 1 - sum((actual_ann - predicted_ann)^2) / sum((actual_ann - mean(actual_ann))^2)
```

Listing 3: R Script for ANN Model

# .4 Climate Data for Zambia between 2013 and 2023

Table 1: Monthly Climate Data (2013)

| Year | Month | Temp. (C) | GHI | Humidity (%) | Precip. (mm) | Wind (m/s) |
|---|---|---|---|---|---|---|
| 2013 | January | 21.8 | 163.73 | 86.00 | 317.00 | 3.70 |
| 2013 | February | 21.0 | 158.22 | 84.87 | 154.00 | 6.70 |
| 2013 | March | 20.5 | 179.34 | 74.59 | 49.20 | 5.30 |
| 2013 | April | 19.6 | 178.65 | 57.50 | 4.80 | – |
| 2013 | May | 17.8 | 170.65 | 46.98 | 0.00 | 5.70 |
| 2013 | June | 16.2 | 149.37 | 45.65 | 0.00 | – |
| 2013 | July | 15.7 | 171.57 | 46.90 | 0.00 | – |
| 2013 | August | 19.4 | 187.12 | 36.73 | 0.00 | – |
| 2013 | September | 23.5 | 207.74 | 31.04 | 0.00 | – |
| 2013 | October | 24.3 | 211.94 | 37.85 | 65.30 | – |
| 2013 | November | 24.1 | 197.47 | 50.09 | 95.70 | – |
| 2013 | December | 22.2 | 192.39 | 72.95 | 170.10 | 4.60 |

Table 2: Monthly Climate Data (2014)

| Year | Month | Temp. (C) | GHI | Humidity (%) | Precip. (mm) | Wind (m/s) |
|---|---|---|---|---|---|---|
| 2014 | January | 21.1 | 181.30 | 85.64 | 133.80 | 3.40 |
| 2014 | February | 20.7 | 144.61 | 87.18 | 141.40 | 4.00 |
| 2014 | March | 21.0 | 178.81 | 81.83 | 51.60 | 5.10 |
| 2014 | April | 19.4 | 156.22 | 69.91 | 93.10 | 5.50 |
| 2014 | May | 18.2 | 169.49 | 57.67 | 0.00 | 5.20 |
| 2014 | June | 16.7 | 150.92 | 48.75 | 0.00 | 5.70 |
| 2014 | July | 16.4 | 169.18 | 46.35 | 0.00 | – |
| 2014 | August | 19.0 | 184.37 | 38.40 | 0.00 | – |
| 2014 | September | 22.3 | 205.79 | 36.02 | 0.00 | 6.10 |
| 2014 | October | 24.9 | 231.31 | 32.92 | 0.00 | 5.10 |
| 2014 | November | 25.7 | 214.05 | 39.52 | 31.30 | 3.80 |
| 2014 | December | 22.6 | 163.14 | 69.88 | 86.10 | – |

Table 3: Monthly Climate Data (2015)

| Year | Month | Temp. (C) | GHI | Humidity (%) | Precip. (mm) | Wind (m/s) |
|---|---|---|---|---|---|---|
| 2015 | January | 20.7 | 172.79 | 84.94 | 109.50 | – |
| 2015 | February | 21.2 | 160.33 | 84.29 | 105.90 | – |
| 2015 | March | 20.8 | 187.58 | 74.12 | 67.70 | – |
| 2015 | April | 19.6 | 139.01 | 75.69 | – | – |
| 2015 | May | 17.9 | 179.71 | 58.09 | 0.00 | – |
| 2015 | June | 16.3 | 160.15 | 49.83 | 0.00 | – |
| 2015 | July | 17.2 | 170.83 | 45.06 | 0.00 | – |
| 2015 | August | 19.5 | 197.48 | 36.67 | 0.00 | – |
| 2015 | September | 22.7 | 211.00 | 32.89 | 0.00 | 3.80 |
| 2015 | October | 25.5 | 222.71 | 32.30 | 0.00 | – |
| 2015 | November | 23.8 | 195.19 | 46.98 | 54.90 | 3.80 |
| 2015 | December | 23.3 | 206.50 | 70.81 | 126.00 | – |

Table 4: Monthly Climate Data (2016)

| Year | Month | Temp. (C) | GHI | Humidity (%) | Precip. (mm) | Wind (m/s) |
|---|---|---|---|---|---|---|
| 2016 | January | 22.2 | 186.34 | 75.66 | 153.10 | – |
| 2016 | February | 22.3 | 169.34 | 78.57 | 163.80 | 3.70 |
| 2016 | March | 21.6 | 158.23 | 81.32 | 123.10 | – |
| 2016 | April | 19.6 | 158.22 | 69.21 | 0.00 | – |
| 2016 | May | 17.1 | 165.16 | 56.01 | 0.00 | 5.10 |
| 2016 | June | 15.4 | 152.74 | 50.40 | 0.00 | – |
| 2016 | July | 16.3 | 155.18 | 47.73 | 0.00 | – |
| 2016 | August | 18.6 | 183.73 | 40.82 | 0.00 | 6.40 |
| 2016 | September | 23.3 | 213.47 | 32.03 | 0.00 | 6.50 |
| 2016 | October | 26.3 | 216.03 | 29.31 | 10.00 | 5.10 |
| 2016 | November | 24.6 | 185.57 | 49.79 | 114.10 | 3.80 |
| 2016 | December | 21.7 | 177.60 | 71.69 | 226.70 | – |

Table 5: Monthly Climate Data (2017)

| Year | Month | Temp. (C) | GHI | Humidity (%) | Precip. (mm) | Wind (m/s) |
|---|---|---|---|---|---|---|
| 2017 | January | 21.0 | 145.67 | 87.61 | 301.10 | 3.00 |
| 2017 | February | 20.9 | 142.73 | 87.88 | 272.90 | 3.40 |
| 2017 | March | 19.7 | 163.62 | 87.24 | 69.90 | 4.90 |
| 2017 | April | 18.9 | 149.24 | 80.57 | 26.00 | 5.90 |
| 2017 | May | 18.4 | 157.78 | 64.31 | 7.00 | – |
| 2017 | June | 16.5 | 148.12 | 55.86 | 0.00 | – |
| 2017 | July | 15.5 | 157.81 | 50.41 | 0.00 | – |
| 2017 | August | 18.3 | 187.06 | 42.34 | 0.00 | – |
| 2017 | September | 22.4 | 210.78 | 34.32 | 0.00 | 5.30 |
| 2017 | October | 25.0 | 199.23 | 37.09 | 12.00 | 4.80 |
| 2017 | November | 22.1 | 159.00 | 59.91 | 38.90 | 4.20 |
| 2017 | December | 21.5 | 181.42 | 72.31 | 144.50 | – |

Table 6: Monthly Climate Data (2018)

| Year | Month | Temp. (C) | GHI | Humidity (%) | Precip. (mm) | Wind (m/s) |
|---|---|---|---|---|---|---|
| 2018 | January | 21.6 | 207.81 | 63.29 | 52.00 | 4.00 |
| 2018 | February | 21.0 | 134.30 | 86.33 | 300.60 | 3.80 |
| 2018 | March | 20.7 | 169.71 | 85.73 | 101.90 | 3.70 |
| 2018 | April | 19.5 | 165.82 | 72.34 | 8.20 | 4.70 |
| 2018 | May | 18.8 | 157.41 | 61.04 | 11.30 | 4.80 |
| 2018 | June | 16.4 | 155.46 | 48.26 | 0.00 | 4.30 |
| 2018 | July | 16.5 | 136.88 | 55.39 | 0.00 | 6.40 |
| 2018 | August | 20.2 | 194.99 | 34.26 | 0.00 | 5.80 |
| 2018 | September | 23.1 | 213.44 | 31.70 | 0.00 | 6.40 |
| 2018 | October | 23.7 | 222.74 | 34.84 | 0.00 | 6.40 |
| 2018 | November | 23.9 | 201.45 | 44.72 | 28.90 | 5.80 |
| 2018 | December | 21.9 | 182.23 | 70.57 | 176.70 | – |

Table 7: Monthly Climate Data (2019)

| Year | Month | Temp. (C) | GHI | Humidity (%) | Precip. (mm) | Wind (m/s) |
|---|---|---|---|---|---|---|
| 2019 | January | 21.3 | 168.60 | 82.15 | 211.80 | 3.40 |
| 2019 | February | 21.8 | 172.92 | 75.65 | 81.70 | 3.90 |
| 2019 | March | 22.8 | 207.63 | 57.87 | 29.60 | 3.60 |
| 2019 | April | 21.0 | 162.90 | 64.51 | 87.20 | – |
| 2019 | May | 18.9 | 172.40 | 54.07 | 0.00 | – |
| 2019 | June | 16.2 | 151.12 | 51.49 | 0.00 | 4.60 |
| 2019 | July | 15.6 | 176.89 | 40.05 | 0.00 | 5.40 |
| 2019 | August | 20.2 | 186.76 | 42.41 | 0.00 | 6.20 |
| 2019 | September | 22.1 | 209.43 | 31.81 | 0.00 | 6.20 |
| 2019 | October | 26.2 | 227.73 | 34.42 | 0.00 | 6.30 |
| 2019 | November | 24.3 | 194.52 | 54.70 | 77.70 | 5.00 |
| 2019 | December | 23.1 | 205.33 | 70.61 | 120.40 | 4.70 |

Table 8: Monthly Climate Data (2020)

| Year | Month | Temp. (C) | GHI | Humidity (%) | Precip. (mm) | Wind (m/s) |
|---|---|---|---|---|---|---|
| 2020 | January | 21.7 | 172.39 | 80.81 | 266.40 | 3.10 |
| 2020 | February | 21.3 | 160.63 | 84.64 | 321.40 | 3.80 |
| 2020 | March | 21.2 | 179.66 | 79.98 | 6.00 | 4.40 |
| 2020 | April | 21.1 | 171.74 | 61.09 | 0.00 | 4.40 |
| 2020 | May | 18.7 | 180.46 | 49.33 | 0.00 | – |
| 2020 | June | 15.6 | 158.24 | 54.44 | 0.00 | – |
| 2020 | July | 15.3 | 157.61 | 53.37 | 0.00 | 6.30 |
| 2020 | August | 19.7 | 193.41 | 38.93 | 0.00 | 5.50 |
| 2020 | September | 22.7 | 209.67 | 38.29 | 0.00 | 6.30 |
| 2020 | October | 25.5 | 223.74 | 37.63 | 5.40 | 6.70 |
| 2020 | November | 25.5 | 200.65 | 48.09 | 114.70 | – |
| 2020 | December | 21.1 | 163.29 | 86.56 | 400.70 | – |

Table 9: Monthly Climate Data (2021)

| Year | Month | Temp. (C) | GHI | Humidity (%) | Precip. (mm) | Wind (m/s) |
|---|---|---|---|---|---|---|
| 2021 | January | 20.9 | 162.71 | 88.19 | 158.30 | – |
| 2021 | February | 21.1 | 150.83 | 88.64 | 204.80 | 3.10 |
| 2021 | March | 20.3 | 182.69 | 84.88 | 93.20 | – |
| 2021 | April | 18.7 | 182.02 | 71.12 | 0.00 | – |
| 2021 | May | 16.8 | 169.21 | 54.65 | 0.00 | 2.30 |
| 2021 | June | 14.8 | 128.19 | 52.97 | 0.00 | – |
| 2021 | July | 15.2 | 155.70 | 50.51 | 0.00 | – |
| 2021 | August | 19.6 | 192.20 | 42.58 | 0.00 | – |
| 2021 | September | 22.3 | 220.55 | 38.75 | 0.00 | 6.20 |
| 2021 | October | 24.8 | 216.03 | 39.35 | 0.40 | – |
| 2021 | November | 22.2 | 199.55 | 56.49 | 72.80 | 5.00 |
| 2021 | December | 24.5 | 214.51 | 60.79 | 64.20 | – |

Table 10: Monthly Climate Data (2022)

| Year | Month | Temp. (C) | GHI | Humidity (%) | Precip. (mm) | Wind (m/s) |
|---|---|---|---|---|---|---|
| 2022 | January | 20.5 | 154.25 | 87.09 | 291.70 | – |
| 2022 | February | 20.9 | 169.15 | 84.51 | 81.60 | 3.60 |
| 2022 | March | 20.8 | 166.82 | 78.48 | 63.80 | – |
| 2022 | April | 19.7 | 156.66 | 73.87 | 24.40 | – |
| 2022 | May | 17.9 | 174.72 | 54.15 | 0.00 | – |
| 2022 | June | 18.4 | 155.48 | 58.60 | 0.00 | – |
| 2022 | July | 15.7 | 149.46 | 57.48 | 0.00 | – |
| 2022 | August | 18.9 | 192.85 | 42.47 | 0.00 | – |
| 2022 | September | 22.3 | 202.46 | 36.83 | 0.00 | – |
| 2022 | October | 25.7 | 229.24 | 31.86 | 0.00 | – |
| 2022 | November | 22.2 | 172.70 | 69.15 | 569.60 | – |
| 2022 | December | 21.1 | 188.62 | 71.97 | 180.90 | – |

Table 11: Monthly Climate Data (2023)

| Year | Month | Temp. (C) | GHI | Humidity (%) | Precip. (mm) | Wind (m/s) |
|---|---|---|---|---|---|---|
| 2023 | January | 20.2 | 153.94 | 86.20 | 409.60 | – |
| 2023 | February | 21.6 | 153.41 | 85.34 | 252.10 | – |
| 2023 | March | 21.1 | 180.38 | 80.39 | 115.70 | 5.10 |
| 2023 | April | 20.7 | 183.61 | 63.02 | 0.00 | – |
| 2023 | May | 19.8 | 181.28 | 47.71 | 0.00 | 0.00 |
| 2023 | June | 18.4 | 158.55 | 39.44 | 0.00 | 0.00 |
| 2023 | July | 16.2 | 169.25 | 49.49 | 0.00 | – |
| 2023 | August | 18.7 | 184.81 | 43.91 | 0.00 | – |
| 2023 | September | 24.1 | 209.20 | 33.79 | 0.00 | 0.00 |
| 2023 | October | 20.5 | 204.05 | 46.44 | 0.00 | – |
| 2023 | November | 25.5 | 219.17 | 48.95 | 66.80 | – |
| 2023 | December | 23.1 | 195.80 | 68.59 | 196.50 | – |

# Bibliography


[1] Government of the Republic of Zambia, "National Energy Policy," Online, 2019, accessed: Jul. 21, 2025. [Online]. Available: https://www.moe.gov.zm/?wpfb_dl=85

[2] Climate Compatible Growth, "Status quo of the energy system and consumption in zambia," https://www.climatecompatiblegrowth.com, 2024.

[3] Zambia Ministry of Energy, "Zambia's integrated resource plan (irp) report," 2023.

[4] Energy Regulation Board, "2024 mid-year statistical bulletin," 2024, unpublished report.

[5] M. Abbott, K. McIntosh, B. Sudbury, J. Meydbray, T. H. Fung, M. U. Khan, Y. Zhang, S. Zou, X. Wang, G. Xing, G. Scardera, and D. Payne, "Annual energy yield analysis of solar cell technology," in *Conf Rec IEEE Photovoltaic Spec Conf*, 2019, pp. 3046–3050.

[6] C. Kapumpu, P. Chisale, N. Kwendakwema, and E. Luwaya, "Predicting global solar radiation on a horizontal surface: A case study for zambia," *Journal of Natural and Applied Sciences*, vol. 4, no. 1, pp. 33–41, 2020.

[7] L.-C. Martin, "Machine learning vs traditional forecasting methods: An application to south african gdp," Stellenbosch University, Department of Economics, Tech. Rep. WP12/2019, 2019. [Online]. Available: https://www.ekon.sun.ac.za/wpapers/2019/wp122019/wp122019.pdf

[8] M. Ajith and M. Martinez-Ramon, "Deep learning algorithms for very short term solar irradiance forecasting: A survey," *Renewable and Sustainable Energy Reviews*, vol. 182, 2023.

[9] O. W. Westbrook, S. M. MacAlpine, and D. A. Bowersox, "Comparison of measured and modeled snow losses for photovoltaic systems in colorado," in *Conf Rec IEEE Photovoltaic Spec Conf*, 2022, pp. 964–966.

[10] J. Weier, "Measuring vegetation (ndvi & evi): Earth observatory feature articles," https://earthobservatory.nasa.gov/features/MeasuringVegetation, 2003.

[11] L. M. A. El-Sayed, D. K. Ibrahim, M. I. Gilany, and A. El'Gharably, "An accurate technique for supervising distance relays during power swing," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 21, no. 3, pp. 1279–1290, 2021.

[12] A. Alzahrani, P. Shamsi, M. Ferdowsi, and C. Dagli, "Solar irradiance forecasting using deep recurrent neural networks," in *2017 IEEE 6th International Conference on Renewable Energy Research and Applications (ICRERA)*, 2017, pp. 988–994.

[13] H. Abubakr, J. C. Vasquez, K. Mahmoud, M. M. F. Darwish, and J. M. Guerrero, "Robust pid-pss design for stability improvment of grid-tied hydroturbine generator," in *Int. Middle East Power Syst. Conf., MEPCON - Proc.*, 2021, pp. 607–612.

[14] A. Aderibole, H. H. Zeineldin, M. S. El-Moursi, J. C.-H. Peng, and M. A. Hosani, "Domain of stability characterization for hybrid microgrids considering different power sharing conditions," *IEEE Transactions on Energy Conversion*, vol. 33, no. 1, pp. 312–323, 2018.

[15] A. A. Shah, P. Leung, Q. Xu, P.-C. Sui, and W. Xing, "Machine learning for flow battery systems," in *Engineering Applications of Computerized Methodologies*. Springer Nature, 2023, vol. 16, pp. 175–284.

[16] S. Badugu and R. Kolikipogu, "Supervised machine learning approach for identification of malicious urls," in *Lecture Notes in Networks and Systems*, ser. LNNS, vol. 119. Springer, 2020, pp. 187–197, scopus.

[17] S. Dutt, S. Chandramouli, and A. K. Das, *Machine Learning*. Pearson India Education Services Pvt. Ltd, 2020.

[18] Y. Zhou, "Advances of machine learning in multi-energy district communities? mechanisms, applications and perspectives," *Energy and AI*, vol. 10, 2022.

[19] K. A. Kumar, B. M. Kumar, A. Veeramuthu, and V. S. Mynavathi, *Unsupervised Machine Learning for Clustering the Infected Leaves Based on the Leaf-Colors*, 2019, vol. 16, pp. 303–312.

[20] R. Abakouy, E. M. En-Naimi, A. El Haddadi, and L. Elaachak, "Machine learning as an efficient tool to support marketing decision-making," in *Lecture Notes in Intelligent Transportation Infrastructures*. Springer Nature, 2020, vol. F1409, pp. 244–258.

[21] T. M. Mitchell, *Machine Learning*, 1st ed., ser. McGraw-Hill Series in Computer Science. New York, NY, USA: McGraw-Hill, 1997.

[22] A. N. Abougreen and C. Chakraborty, "Applications of machine learning and internet of things in agriculture," in *Green Technology Innovation for Sustainable Smart Societies: Post Pandemic Era*. Springer International Publishing, 2021, pp. 257–279.

[23] M. Devgan, G. Malik, and D. K. Sharma, "Semi-supervised learning," in *Machine Learning and Big Data: Concepts, Algorithms, Tools and Applications*. Wiley, 2020, ch. 10, pp. 251–280, scopus.

[24] Y. S. Afridi, L. Hassan, and K. Ahmad, "Machine learning applications for renewable energy systems," in *EAI/Springer Innovations in Communications and Computing*. Springer, 2023, vol. F665, pp. 79–104.

[25] S. Aboukadri, A. Ouaddah, and A. Mezrioui, "Major role of ai, machine learning, and deep learning in identity and access management: Challenges and state of the art," in *Lecture Notes in Data Engineering and Communications Technologies*. Springer, 2023, vol. 152, pp. 50–64.

[26] C. A. Gueymard, *Solar Radiation Resource: Measurement, Modeling, and Methods*. Elsevier, 2022, vol. 1, pp. 176–212.

[27] M. Aakroum, A. Ahogho, A. Aaqir, and A. A. Ahajjam, "Deep learning for inferring the surface solar irradiance from sky imagery," in *Proceedings of the International Renewable and Sustainable Energy Conference (IRSEC)*, Y. Zaz and M. Essaaidi, Eds. IEEE, 2018.

[28] T. Ahmad, D. Zhang, and C. Huang, "A comprehensive review of solar irradiance forecasting models," *Renewable and Sustainable Energy Reviews*, vol. 52, pp. 975–992, 2020.

[29] V. Akshay, "Solar irradiance calculation guide," Online; Republic of Solar, 2022, accessed: 2024. [Online]. Available: https://arka360.com/ros/solar-irradiance-calculation/

[30] R. Aler, I. M. Galván, J. A. Ruiz-Arias, and C. A. Gueymard, "Improving the separation of direct and diffuse solar radiation components using gradient boosting," *Solar Energy*, vol. 150, pp. 558–569, 2017.

[31] M. E. Haque *et al.*, "Cost effective alternative of pyranometer: Solar radiation prediction using artificial intelligence," in *ICRPSET 2022*, 2022.

[32] L. Ait Mouloud, A. Kheldoun, A. Deboucha, and S. Mekhilef, "Explainable forecasting of global horizontal irradiance using temporal fusion transformer," *Journal of Renewable and Sustainable Energy*, vol. 15, no. 5, 2023.

[33] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 2nd ed. Melbourne, Australia: OTexts, 2018. [Online]. Available: https://otexts.com/fpp2/

[34] S. R. Walk, "Quantitative technology forecasting techniques," *Technological Change*, vol. 24, 2012.

[35] J. S. Armstrong, "Selecting forecasting methods," in *Principles of forecasting: A handbook for researchers and practitioners*. Springer, 2001, pp. 365–386.

[36] J. C. Chambers and S. K. Mullick, "Forecasting for planning: Qualitative techniques," *Planning Review*, vol. 3, no. 5, pp. 13–27, 1975.

[37] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "The m4 competition: Results, findings, conclusion, and way forward," *International Journal of Forecasting*, vol. 35, no. 1, pp. 17–25, 2019.

[38] P. R. Winters, "Forecasting sales by exponentially weighted moving averages," *Management Science*, vol. 6, no. 3, pp. 324–342, 1960.

[39] T. Cebecauer, D. Chrkavy, N. Suriova, B. Schnierer, J. Betak, A. Skoczek, and M. Suri, "Solar resource and photovoltaic power potential of [region/country, if known]," Solargis s.r.o., Mytna 48, 811 07 Bratislava, Slovakia, Tech. Rep., 2018, reference No. 128-08/2018. [Online]. Available: http://solargis.com

[40] G. Box, "Box and jenkins: time series analysis, forecasting and control," in *A Very British Affair: Six Britons and the Development of Time Series Analysis During the 20th Century*. Springer, 2013, pp. 161–215.

[41] J. J. Heckman and E. J. Vytlacil, "Econometric evaluation of social programs, part i: Causal models, structural models and econometric policy evaluation," *Handbook of econometrics*, vol. 6, pp. 4779–4874, 2007.

[42] D. R. Myers, "Direct normal radiation," in *Handbook of Concentrator Photovoltaic Technology*. Wiley, 2016, ch. 1, pp. 1–58, scopus.

[43] J. Brownlee, *Introduction to Time Series Forecasting with Python: How to Prepare Data and Develop Models to Predict the Future*. Machine Learning Mastery, 2017, online Book. [Online]. Available: https://machinelearningmastery.com/time-series-forecasting/

[44] C. Zhang, L. Yan, and J. Shi, "Performance prediction of a supercritical $CO_2$ brayton cycle integrated with wind farm-based molten salt energy storage: Artificial intelligence (AI) approach," *Case Studies in Thermal Engineering*, vol. 51, p. 103533, 2023, scopus.

[45] Y. Zou and X. Yang, "Machine learning in energy forecasting: Applications and improvements," *Journal of Energy Engineering*, vol. 148, no. 5, pp. 123–134, 2022.

[46] T. Ahmad, D. Zhang, and C. Huang, "A comprehensive review of solar irradiance forecasting models," *Renewable and Sustainable Energy Reviews*, vol. 52, pp. 975–992, 2020, review Article.

[47] B. Müller and F. Scherer, "Strategic long-term planning for renewable energy," *Energy Policy*, vol. 113, pp. 252–262, 2018.

[48] J. W. Hall and S. T. Buckland, "Long-term predictions and scenario analysis for environmental change," *Environmental Modelling & Software*, vol. 89, pp. 1–6, 2017.

[49] IPCC, "Special report on climate change and land," Intergovernmental Panel on Climate Change, Special Report, 2019. [Online]. Available: https://www.ipcc.ch/srccl/

[50] C. Kapumpu, "Predicting the global solar radiation on a horizontal surface: A case study for zambia," Ph.D. dissertation, The University of Zambia, 2015.

[51] S. Srivastava and S. Lessmann, "A comparative study of lstm neural networks in forecasting day-ahead global horizontal irradiance with satellite data," *Solar Energy*, vol. 162, pp. 232–247, 2018.

[52] P. Lara-Benitez, M. Carranza-Garcia, J. M. Luna-Romera, and J. C. Riquelme, "Short-term solar irradiance forecasting in streaming with deep learning," *Neurocomputing*, vol. 546, 2023.

[53] A. F. Kamga and N. Djongyang, "Machine learning techniques for solar energy prediction in cameroon: A case study on short-term forecasting," *Journal of Solar Energy*, vol. 34, no. 4, pp. 450–463, 2019.

[54] S. El Bakali, O. Hamid, and S. Gheouany, "Day-ahead seasonal solar radiation prediction, combining vmd and stack algorithms," *Clean Energy*, vol. 7, no. 4, pp. 911–925, 2023.

[55] K. Ullah, M. Ahsan, S. M. Hasanat, M. Haris, H. Yousaf, S. F. Raza, R. Tandon, S. Abid, and Z. Ullah, "Short-term load forecasting: A comprehensive review and simulation study with cnn-lstm hybrids approach," *IEEE Access*, 2024.

[56] E. Perez, J. Perez, J. Segarra-Tamarit, and H. Beltran, "A deep learning model for intra-day forecasting of solar irradiance using satellite-based estimations in the vicinity of a pv power plant," *Solar Energy*, vol. 218, pp. 652–660, 2021.

[57] P. Mpfumali, C. Sigauke, A. Bere, and S. Mulaudzi, "Day ahead hourly global horizontal irradiance forecasting-application to south african data," *Energies*, vol. 12, no. 18, 2019.

[58] A. H. Nielsen, A. Iosifidis, and H. Karstoft, "Irradiancenet: Spatiotemporal deep learning model for satellite-derived solar irradiance short-term forecasting," *Solar Energy*, vol. 228, pp. 659–669, 2021.

[59] J. Lago, K. De Brabandere, F. De Ridder, and B. De Schutter, "Short-term forecasting of solar irradiance without local telemetry: A generalized model using satellite data," *Solar Energy*, vol. 173, pp. 566–577, 2018.

[60] N. E. Michael, S. Hasan, A. Al-Durra, and M. Mishra, "Short-term solar irradiance forecasting based on a novel bayesian optimized deep long short-term memory neural network," *Applied Energy*, vol. 324, 2022.

[61] R. Wang, "Application of machine learning in prediction of urban heat island," in *Advances in 21st Century Human Settlements*. Springer, 2023, pp. 171–206.

[62] D. C. Montgomery, C. L. Jennings, and M. Kulahci, *Introduction to time series analysis and forecasting*. John Wiley & Sons, 2015.

[63] F. C. S. Quevedo, "A comparison of machine learning and traditional demand forecasting methods," Master's thesis, Clemson University, 2020.

[64] J. Thaker, R. Hoeller, and M. Kapasi, "Short-term solar irradiance prediction with a hybrid ensemble model using eumetsat satellite images," *Energies*, vol. 17, no. 2, 2024.

[65] H. Lee and B. Lee, "Bayesian deep learning-based confidence-aware solar irradiance forecasting system," in *Proceedings of Electronics Telecommunications Research Institute - Korea (ETRI)*. IEEE, 2018, pp. 1233–1238.

[66] H. Panamtash, S. Mahdavi, Q. Z. Sun, G.-J. Qi, H. Liu, and A. Dimitrovski, "Very short-term solar power forecasting using a frequency incorporated deep learning model," *IEEE Open Access Journal of Power and Energy*, vol. 10, pp. 517–527, 2023.