SHORT- AND LONG-TERM FORECASTING OF WIND SPEED IN LIMPOPO PROVINCE USING MACHINE LEARNING ALGORITHM AND EXTREME VALUE THEORY

by

KGOTHATSO MAKUBYANE

MINI DISSERTATION

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

E-SCIENCE

in the

FACULTY OF SCIENCE AND AGRICULTURE
SCHOOL OF MATHEMATICAL AND COMPUTER SCIENCES

at the

UNIVERSITY OF LIMPOPO

SUPERVISOR: PROF. D MAPOSA

FEBRUARY 2024

Declaration

I, **Kgothatso Makubyane**, therefore acknowledge that the mini dissertation titled ”Short- and long-term forecasting of wind speed in Limpopo Province using machine learning algorithm and extreme value theory” is my original work. All material from other sources including any works created by other persons or organisation that was used in this research study was appropriately attributed and referenced. I additionally testify that this research study has never been submitted by anybody from another university.

Signature:........*KM*............Date:......03 February 2024..........
Makubyane, K.

Copyright © 2023 University of Limpopo
All rights reserved

Abstract

Numerous studies have applied Extreme Value Theory (EVT) to model environmental variables like wind speed, rainfall and temperature. Recently, academic focus has shifted to machine learning algorithms for the same variables. This research study demonstrates the practical use of EVT and machine learning techniques for modelling wind speed in the Limpopo Province, with the primary goal of assessing wind power generation reliability. The data used in this research study is obtained from National Aeronautics and Space Administration (NASA), spanning the time period from 2016 to 2022. The Vanilla Long Short-Term Memory (LSTM) network exhibited remarkable accuracy, achieving 86% training and 89% testing accuracy. Additionally, Generalised Extreme Value Distribution (GEVD) for block sizes (1 to 5) revealed $GEVD_{m=2}$ as the most suitable model based on low Akaike information criteria (AIC) and Bayesian information criteria (BIC) values. The model highlighted a rare event with a 300-year return period, indicating a wind speed of 22.893 meters. This study provides valuable insights for careful power planning, economic strategy and advancement in civilisation in South Africa, with implications for future energy planning and policy decisions in the region.

Keywords: Akaike information criteria, Bayesian information criteria, Extreme Value Theory, Generalised Extreme Value Distribution, Long Short-Term Memory, National Aeronautics and Space Administration, and Wind Power Generation

Dedication

I dedicate this research study to my daughter Thatego Mahlangu and my parents, Doris Kanyane Makubyane, Thabo Chadwick Mogashoa, and to all my siblings, whose loyal love, affection, and inspiration have been the base foundation behind my academic journey.

Acknowledgments

I am grateful to the most high God who is above us all for the wisdom, fortitude and bravery he has given me throughout my academic path. My deepest thanks and gratitude goes out to my supervisor, Prof. D. Maposa, for his patience, encouragement, insightful remarks and ideas, not to mention the assistance provided by Dr. Helen Robertson from the University of Witwatersrand. Finally, I would want to thank the Department of Statistics and Operations Research, the academic staff who taught me from the undergraduate to postgraduate levels and the entire Team NEPTTP for granting me the funding to further my education.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and background

1.1 Introduction

An outstanding advancement in technology revolution carried by the wind power generation is happening within Asian and European countries (Zhang et al., 2022). This approach can help South Africa deal with its ongoing electricity problems, opening the way for a better and more successful future for its people, businesses and economy. However, to maintain wind power stability, wind power generation requires careful planning and execution, as well as accurate wind speed prediction in the implementation region. This implies that the core component of putting this idea into practice is wind speed forecasting.

The complexity of atmospheric processes that affect wind speed is one of the obstacles in forecasting wind speed. Factors including humidity, surface roughness, temperature ($^{\circ}C$) and atmospheric pressure gradients all have an impact on wind speed prediction (Ahmed et al., 2022; Antonini and Caldeira, 2021).

In order to provide reliable forecasts, forecasting models must take into consideration these factors and their interactions. Additionally, warm air is preferred by wind turbines in order to pull more kinetic energy from the environment (Fischereit et al., 2022; Jensen and Skelton, 2018). Due to scarcity of the literature pertaining to semi-developed provinces, the study aims to showcase the potential to produce electricity through wind in the province of Limpopo.

In recent decades, Wang et al. (2020) stated that the modern machine learning algorithms have significantly improved the prediction precision of meteorological features. More accurate predictions have been made achievable by the use of high-resolution weather models in machine learning, deep learning and Extreme Value Theory (EVT), together with enhanced data integration techniques. In this research study, a classical statistical model called Generalized Extreme Value Distribution (GEVD) is explored, which combines the heavy tailed Fréchet distribution, light tailed of a Weibull distribution and Gumbel distribution, yielding a unified distribution (Coles et al., 2001). This research study also discusses the Vanilla Long Short-Term Memory (LSTM) network which is an advanced form of Recurrent Neural Network (RNN). The advantages of these models include their great adaptability and ability to adapt to variety of datasets, including time series and speech translation.

Wind power generation can significantly boost local economic development, provide a cost-effective source of energy and lower the unemployment rate since wind turbines need qualified employees to construct and maintain them. This research study asserts the importance of continued financial support from government authorities towards the development and enhancement of wind power infrastructure and technologies, to raise awareness among government officials and the Electricity Supply Commission of South Africa (Eskom) to understand the efficiency and advantages of wind power generation.

1.2 Background

This research study employs both machine learning algorithms and Extreme Value Theory (EVT) methods. However, the primary focus is on a robust machine learning technique and comprehensive EVT statistical approaches. The Vanilla Long Short-Term Memory (LSTM) network, a modification of Recurrent Neural Networks (RNN), is chosen due to its ability to address the vanishing information problem in distant history. Hochreiter and Schmidhuber (1997) developed this technique. The capacity of this algorithm is to recall or forget certain information at each time step. This is accomplished by using input gate, forget gate and output gate which regulate the information flow into and out of the LSTM cell (Geng et al., 2020). This selection is justified to enhance the depth of our modelling approach.

On the one hand, numerous researchers exploited the (GEVD), which combines the Fréchet, Weibull and Gumbel parametric distributions, for applying EVT to analyze wind, temperature and rainfall (Beirlant et al., 2006; Coles et al., 2001). This motivated us to employ the classical approach, specifically the GEVD, for the examination of wind speed data. Maposa et al. (2021) estimated the temperature return levels for 10, 20, 50 and 100 years. Both the GEVD and Generalized Pareto Distribution (GPD) appeared to suit the data of maximum temperature returns well, supported by the diagnostic plots. The best model, however, was selected based on the lowest Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) values. In neural networks, the optimality of the model is examined by measuring the model accuracy on testing data after the model has been trained.

1.3 Problem statement

According to the Council for Scientific and Industrial Research (CSIR), South Africa experienced the worst load shedding during the years: 2019–2022 (CSIR, 2023). This condition appears to be growing worse and worse with time (Mbandlwa, 2023; Banderker, 2022). However, wind energy generation can provide a solution to this obstacle (Madiba et al., 2022; Daniel et al., 2020). The combustion of fossil fuels is one of Africa's traditional methods of energy generation, which causes one of the most catastrophic events known as global warming (Al-Ghussain, 2019; Sivaramanan, 2015). This results in unfavourable meteorological conditions, such as elevated earth temperature, air pollution and flooding (Perera and Nadeau, 2022). The aforementioned conditions are significant threat to the economic sector, agricultural sector and health sector.

In the economic sector, extreme weather conditions and natural catastrophes can lead to supply chain disruptions, infrastructure damage and business interruptions (Botzen et al., 2019). In the agricultural sector, high temperatures and flooding may harm crops, diminish yields and raise the danger of pests and diseases (Diffenbaugh and Burke, 2019; Dottori et al., 2018; Haines et al., 2006). In the area of health, production of energy from the burning of fossil fuels has detrimental effects on human health, increasing the risk of respiratory and cardiovascular conditions (Donaghy et al., 2023). Wind power generation can mitigate these challenges, ensuring a more reliable food supply in the country and reducing death rates from cardiovascular diseases.

World Health Organization (2019) highlighted how the increased frequency of traffic accidents is a result of the extreme weather brought on by the use of fossil fuels for electricity generation. Fossil fuel combustion is a significant source of airborne fine particulate matter denoted by a scientific name, $PM_{2.5}$ and a significant factor in the burden of disease and mortality experienced worldwide

(Deng et al., 2014). According to Cohen et al. (2017), about 4.2 million people worldwide died as a result of exposure to $PM_{2.5}$. This research study highlights wind power as a crucial solution in the face of the climate crisis. This not only provides renewable energy solution, but also offers environmental benefits and economic growth. Embracing wind power can create a cleaner, prosperous and harmonious sustainable future for South Africa.

1.4 Rationale

In the fields of EVT and machine learning, limited studies have been conducted on forecasting wind speed for energy generation. Renewable energy production is one crucial factor for the economic progress of each nation (Saulat et al., 2021). This actively contributes to economic progress by optimizing energy production strategies, reducing dependency on traditional energy sources, fostering job creation, encouraging investments in renewable energy infrastructure, driving economic growth and environmental sustainability simultaneously. In recent decades, several scholars suggested time series prediction techniques, such as the Autoregressive Integrated Moving Average (ARIMA) model and the Seasonal Autoregressive Integrated Moving Average (SARIMA) on forecasting wind speed (Elsaraiti and Merabet, 2021; Al Dhaheri et al., 2017). These methods require a lot of data and are less precise compared to Artificial Neural Network (ANN) and Support Vector Regression (SVR) with superior predicting performance (Shao et al., 2021; Bokde et al., 2019; Wang et al., 2016). In a related study, Mutavhatsindi et al. (2020) explored multiple machine learning algorithms, including Vanilla LSTM network, SVR and Feed Forward Neural Network (FFNN), to estimate solar irradiance in South Africa. Among these techniques, FFNN emerged as the best forecasting model, displaying the lowest Root Mean Square Error (RMSE) and Mean Absolute Error (MAE).

In South Africa, a quick transition to a cleaner and more sustainable energy source, like as wind is required. A coordinated effort from governments, corporations and individuals is needed to promote and invest in renewable alternatives, particularly wind power generation. Furthermore, in order to properly develop advanced wind power turbines, predicting extreme wind events is crucial. Through the application of EVT using the GEVD we can match the wind speed distribution that allows for a more accurate estimation of the frequency of high wind occurrences (Morgan et al., 2009). In the previous studies, similar and comparable EVT techniques have been employed in wind speed modelling (Oktaviarina and Sofro, 2019; Morgan et al., 2009; Brabson and Palutikof, 2000).

Castillo et al. (2005) described the blended GEVD (bGEVD) as a versatile and effective technique for simulating distribution tail behaviour. The idea of blending was described in Castillo et al. (2005), along with how it might help with severe event modelling. Also, they provided a breakdown explanation of the bGEVD model parameters. Several literature indicated that it is not sufficient to only fit a particular technique, test for goodness-of-fit is important to evaluate how well a particular probability distribution fits a collection of observed data (Rolke and Gongora, 2021; Zeng et al., 2015).

South Africa is experiencing large-scale power outages (load shedding of stages 5 and 6 for which in a day electricity goes off three times a day for 1 to 4 hours each time). Hence, this research study investigates the possibility to enhance economic and social well-being of South Africa by providing statistical methodologies demonstrating the potential of wind energy as a source of power using statistical EVT and machine learning techniques.

1.5 Aim and objectives

1.5.1 Aim

The research aims to forecast short-term wind speed using the Vanilla Long Short-Term Memory (LSTM) algorithm and long-term wind speed using the Generalized Extreme Value Distribution (GEVD).

1.5.2 Objectives

The objectives of the study are to:

1. Fit the Vanilla LSTM algorithm and GEVD to the wind speed dataset.

2. Assess the predictive performance of the Vanilla LSTM and forecast short-term wind speed using the machine learning algorithm.

3. Assess the goodness-of-fit of the GEVD in long-term modelling of extreme wind speed.

4. Conclude on the optimal block size for modelling extreme events using GEVD, highlighting the most effective size based on the study findings.

5. Estimate the return levels of extreme wind speed and their corresponding return periods using GEVD.

1.6 Contribution of the study

It is becoming more and more common and practical to produce power using renewable resources to meet global energy demands. Yet, the focus of this study is on the role that wind energy plays in producing electricity. Many and continuous scientific advancements have been made in the field of producing power from the wind. The fact that it is free and causes less damage to the environment is one of its biggest advantages to the economic development of any nation. Specifically, wind power offers benefits such as reduced greenhouse gas emissions and minimal air pollution, making it a sustainable and environmentally friendly solution for powering our energy needs.

This research study will advance the field of renewable energies research and assist in reducing the damage done by the combustion of fossil fuels. Wind power generation can significantly boost local economic development, provide a cost-effective source of energy, and lower the unemployment rate since wind turbines need qualified employees to construct and maintain them. This research study asserts the importance of continued financial support from government authorities towards the development and enhancement of wind power infrastructure and technologies.

1.7 Structure of the dissertation

The outline of this research study is divided into, five chapters. **Chapter 2** explores the relevant literature reviews, their findings on the subject matter and research gap identification. **Chapter 3** provides the analytical techniques employed for forecasting wind speed, area of study, data sources and assessment metrics. **Chapter 4** discusses the findings from the data analysis for each technique, interpretation and discussion of the overall results. Finally, **Chapter 5** summarise the major findings, conclusion and recommendations.

Chapter 2

Literature review

2.1 Introduction

This chapter provides a comprehensive review of existing research conducted in South Africa, focusing on renewable energy resources, expanding to different countries and regions on renewable energy power generation. It highlights the importance of renewable resources in addressing global warming and economic development. Additionally, the chapter evaluates the current body of literature, identifies gaps and constraints in the existing literature.

2.2 Overview of literature in South Africa

In recent studies, Daniel et al. (2020) explored different forecasting methods for wind speed, with a focus on point and interval forecasting in Vredandale, South Africa. They used three methods for point forecasting and two methods for combining forecasts, using a dataset from Vredendal in South Africa.

One notable aspect of their study is the use of Generalized Additive Models (GAMs) for wind speed forecasting, which has not been extensively studied in prior literature. The GAMs are known for their interpretability and flexibility in modelling relationships between variables. The researchers evaluated the accuracy of these models using various metrics, such as coverage width-based criteria and prediction interval normalised average width. The results indicated that certain methods perform better than others in terms of accuracy and performance. This study contributed to the field of wind speed forecasting by evaluating different methods and demonstrating their effectiveness in various scenarios.

In the same vein, Mutavhatsindi et al. (2020) explored different machine learning algorithms, including FFNN, Vanilla LSTM network, SVR and PCA as a benchmark model. A benchmark model is a baseline model that serves as a benchmark for evaluating the effectiveness of other models or methodologies. These models were trained and tested using historical weather and solar irradiance data from various locations in South Africa. The study identified the FFNN model as the most suitable forecasting model, which resulted in the best forecast accuracy when measured by MAE and RMSE. Their findings offered valuable insights for the development of efficient and effective solar energy systems in South Africa. Furthermore, this serves as a foundation for future research on utilising machine learning techniques for wind speed prediction in electricity generation.

Diriba et al. (2015) examined the peaks of the yearly and daily maximum wind speeds in Port Elizabeth, South Africa, using the dependence impact to extreme value distributions. The maximum likelihood and Markov chain Monte Carlo approach with the Metropolis-Hastings algorithm were used to estimate the parameters of EVT models (Coles et al., 2001). The outcomes of the Bayesian study demonstrate that the priors selected to create the informative priors may have an impact on posterior inference. In contrast to the frequentest method, the Bayesian approach offered a suitable estimating procedure that accounted for parameter and return level uncertainty. Sigauke and Bere (2017) also studied the impact of choosing the best technique for parameter estimation.

In a separate study, a modelling framework was developed by Sikhwari et al. (2022) using the data on the maximum measures of rainfall that were recorded in the province of Limpopo from 1960 to 2020. A Generalized Pareto Distribution (GPD) and the $GEVD_r$ modelling technique were both applied. Since selecting a substantial r is important in the r-largest order statistics technique, the focus was only on $r < 9$. After the selected appropriate model for the data, $GEVD_{r=8}$, the 50-year return level was predicted to be 368 mm, which indicates a chance of 0.02 surpassing 368 mm in 50 years in the Limpopo Province.

Bhagwandin (2017) compared the univariate and multivariate extreme value theory models to analyse climatic data in Western Cape province, South Africa. Data obtained from five weather stations, Cape Town International Airport, Langebaanweg, George Airport, Vredendal and Plettenberg Bay worth of data were gathered for the analysis beginning in 1965. The study modelled several weather variables including maximum rainfall, maximum wind speed and maximum temperature using block maxima, threshold excess and point process techniques. The results indicated that the block maxima technique was inferior in modelling the weather variables, due to ignoring important observations. The threshold excess and point process techniques, on the other hand, performed better when modelling weather extremes. However, there was slight correlation observed between wind speed, rainfall and temperature, suggesting that future research may need to apply machine learning algorithms.

2.3 Worldwide overview of literature

Bhaskaran et al. (2023) investigated the estimation of extreme wind and waves, demonstrating their relevance to offshore wind energy. The statistical techniques used include the block-maxima and peaks-over-threshold (POT) method, using distribution models such as Gumbel, GEVD and GPD. The mean residual life approach was used to determine a threshold of 4.31 m. The findings revealed variability in extreme values among different sites, highlighting the accuracy of the block-maxima approach. In terms of future research recommendations, the emphasis was on critically exploring alternative threshold determination methods, especially in optimising the POT approach.

The accurate estimation of parameters is essential to ensure dependable values for the prediction of future extreme events (Coles et al., 2001). Abdulali et al. (2022) examined three estimate techniques for estimating parameter values based on simulated observations from the GEVD, namely the method of moments (MOMs), maximum likelihood estimator (MLE) and maximum product of spacing (MPS). The results indicated that MLE outperformed the other techniques in terms of mean square errors, while maintaining similar goodness-of-fit statistics. In particular, for both the location and scale parameters, MLE exhibited the lowest mean square error, followed by MPS, with MOMs performing the least effectively. In summary, both MLE and MPS showed nearly identical performance, while MOMs lagged behind in terms of accuracy.

Arora et al. (2018) presented a comparison of statistical and machine learning approaches, including ARIMA, SVR, and LSTM for two agricultural zones in India. The performance of the proposed models were assessed on the basis of several error measures, such as MAE, MSE and RMSE. According to their findings, each method effectiveness differed across various wind farm zones. Forecasting accuracy was significantly impacted by the choice of model parameters and the selection of suitable kernels. In addition, the study recommended investigating hybrid strategies combining statistical and machine learning techniques for greater prediction accuracy.

The present study focuses on the application of the Vanilla LSTM networks for short-term wind speed forecasting. This choice is based on the Vanilla LSTM network ability to handle long-term dependencies and adapt to dynamic situations. By using this technique, the machine learning model developed will be able to generalise well to any relevant dataset. In a related study, Geng et al. (2020) simulated hourly wind speed data in Xinjiang province, China. They used various meteorological variables such as temperature, humidity, air pressure and wind speed. Principal component analysis was employed to select significant features for short-term wind speed forecasting. These selected features were then inputted into the LSTM network, which demonstrated its effectiveness in modelling wind speed.

In a recent study, Malakouti et al. (2022) proposed a new approach that combines the power of Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) with other machine learning algorithms to improve wind power generation forecasts, based on Texas wind turbine data collected hourly. This approach outperformed SVR, Random Forest (RF) and Gradient Boosting (GB) in terms of accuracy. The experimental results demonstrated that the CNN-LSTM hybrid approach surpassed all other machine learning-based techniques and existing forecasting methods. As a result, this hybrid approach showed a great promise for practical applications in wind speed prediction.

2.4 Importance of renewable resources

Renewable energies are becoming increasingly popular as sources of clean, abundant, and competitive energy. Unlike fossil fuels, they have the advantage of being diverse and available anywhere on the planet. One of the key benefits of renewable energies is that they do not produce greenhouse gases, which are responsible for climate change. Another advantage is that renewable energies are cost-free, while the cost of fossil fuels continues to rise (Panwar et al., 2011). The growth of clean energies is evident in the latest statistics from the International Energy Agency (IEA). According to their forecasts, the share of renewable energy generation supply is expected to increase from 28.7% in 2021 to 43% in 2030 (IEA, 2020). Furthermore, renewables will also contribute to two-thirds of the increase in electricity demand during this period. This growth will primarily be driven by wind energy, solar energy and geothermal energy.

Overall, renewable energies offer a promising solution to our energy needs in South Africa. They provide a sustainable alternative to fossil fuels, offering a cleaner and more abundant source of power. As we continue to invest in and develop renewable technologies, we can expect to see a significant shift towards cleaner energy production in the coming years.

Renewable energy offers numerous economic benefits, particularly in terms of job creation (Arent et al., 2011). Unlike fossil fuels, the renewable energy generation industry is more labour intensive. For example, solar panels require humans to install them, while wind farms need technicians for maintenance. As a result, renewable energy generates more jobs per unit of electricity generated compared to fossil fuels. In the United States alone, renewable energy already supports thousands of jobs (Carley et al., 2021). In 2016, the wind energy industry directly employed over 100,000 full-time equivalent employees in various roles such as manufacturing, project development, construction and turbine installation, operations and maintenance, transportation and logistics, and financial, legal, and consulting services (AWEA, 2017). The use of renewable energies not only contributes to a cleaner environment, but also stimulates economic growth by providing employment opportunities across different sectors.

2.5 Identification of existing research gaps

Research gap identification is crucial to ensure that researchers are not duplicating existing studies and are focusing on areas that require further exploration. In the field of EVT and machine learning, there exists a considerable body of literature. However, there is a significant methodological gap in utilising modern machine learning and deep learning algorithms to model hydrological features and capture complex patterns due to their complex implementation. For instance, Sikhwari et al. (2022); Mashishi et al. (2020); Agilan and Umamahesh (2017) conducted a study on rainfall modelling in selected regions using EVT methods. In other words, more than 70% of studies applied EVT and time series techniques in modelling rainfall, wind speed and temperature.

As we find ourselves in a fast-paced technological environment of the 21st century, this research study aims to address this gap by employing a robust machine learning algorithm called Vanilla LSTM network, renowned for its superior predictive accuracy. Furthermore, there is a geographic gap as many researchers predominantly focus on developed countries and financial hub cities worldwide. Soman et al. (2010) examined wind power and wind speed forecasting techniques for wind power generation in North America across various time periods. This present study stands out by focusing on the semi-developed province of Limpopo in South Africa. The major goal is to aid rural areas in South Africa to generate electricity through renewable resources.

Moreover, it is worth noting that while prior research has focused on specific geographic regions, our study intends to contribute to the broader understanding of hydrological modelling by evaluating the effectiveness of the Vanilla LSTM network in a different regional context. By doing so, we aim to provide valuable insights and methodologies that can be adapted to various geographical areas, thereby promoting the scalability and applicability of our findings in diverse settings.

This research study also adds to the limited number of literature sources in South Africa and globally by integrating the modelling of long-term extreme wind events and short-term events within a single study. Lastly, by combining the essential concepts of machine learning and extreme value theory, this research contributes to various fields such as meteorology, finance, engineering and beyond.

2.6 Summary of the chapter

This chapter reviewed the literature undertaken worldwide by other researchers in the fields of EVT and machine learning, as well as the literature undertaken in South Africa on wind speed and other climatic and meteorological features. This chapter also examined the gaps in the body of literature that need to be filled and highlights them for future investigation. In previous decades literature, GEVD and GPD were used to predict climate data. Similar techniques were used by Chapman et al. (2023); Mashishi et al. (2020) to model maximum rainfall for selected regions. The current study will combine advanced EVT and machine learning methods in a single study to predict both short and long-term wind speed for electricity generation development in South Africa.

Chapter 3

Methodology

3.1 Introduction

This chapter describes the analytical techniques employed in this research study which involves the use of supervised machine learning model, specifically the Vanilla Long Short-Term Memory (LSTM) algorithm and the traditional Extreme Value Theory (EVT) method Generalised Extreme Value Distribution (GEVD), to forecast wind speed in Limpopo Province. The parameters for the GEVD will be estimated using the maximum likelihood estimation (MLE) technique. We deployed the three forecasting evaluation metric methods, namely Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) to evaluate the accuracy of the Vanilla LSTM algorithm. The Akaike information criteria (AIC) and Bayesian information criteria (BIC) goodness of fit method will be used to estimate the prediction error of the GEVD. Testing for stationarity and trend analysis using Theil-Sen estimator. Finally, the study estimate return levels corresponding to their return periods for long-term wind speed future events outcomes.

3.2 Research methodology

3.2.1 Data source and study area

The study will make use of secondary data from the National Aeronautics and Space Administration (NASA) government. Polokwane, our focus station, is situated at latitude $2354'16.16'S$ and longitude $2928'7.86'E$. The dataset includes a variety of meteorological characteristics, including surface pressure, wind direction and earth surface temperature. Beginning on 1 January 2016 and continuing until 31 December 2022, all features are recorded daily during midnight. Below is the wind heatmap representing the wind patterns in the northern and eastern provinces of South Africa.

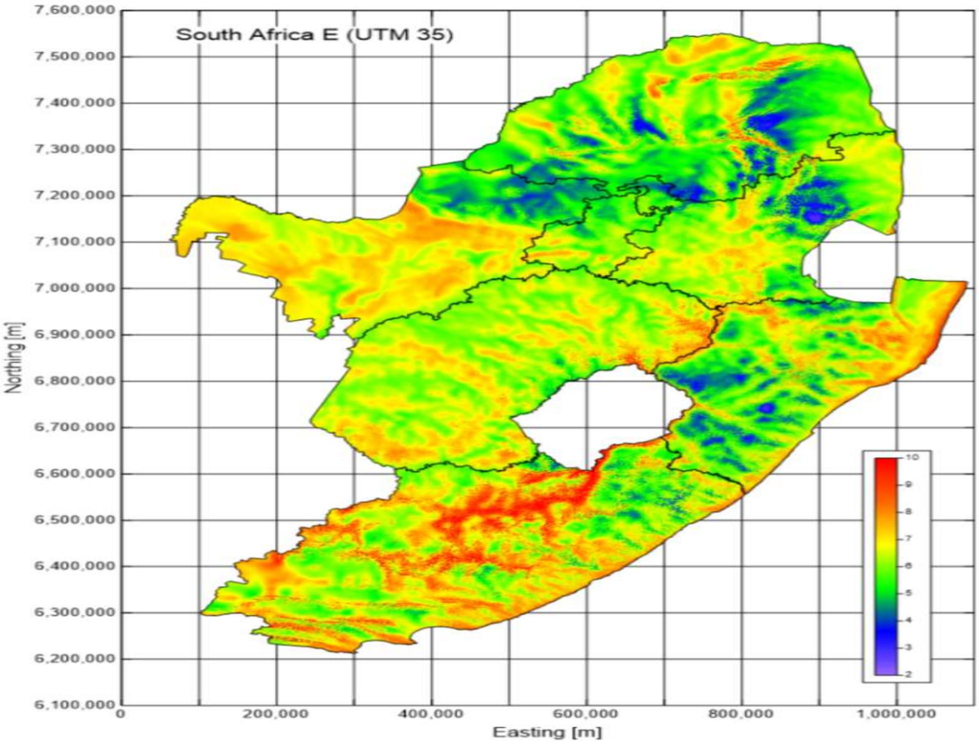


Figure 3.1: Northern and Eastern provinces of South Africa on the Wind Atlas Map. Source: `http://wasadata.csir.co.za/wasa1/WASAData`

3.2.2 Tests for stationarity

Augmented Dickey-Fuller test

In this section, we explain the formal assessment of stationarity using the Augmented Dickey-Fuller (ADF) test. The ADF test is a statistical method employed to determine whether a time series dataset exhibits stationarity (Dickey and Fuller, 1979). Stationarity is a crucial concept in time series analysis as it assumes that the statistical properties of the data remain constant over time. In our research study, we applied the ADF test to assess the stationarity of the daily wind speed data in Polokwane. The formal representation of the ADF hypothesis testing is as follows:

H_0 : Daily wind speed data of Polokwane is non-stationary,
H_1 : Daily wind speed data of Polokwane is stationary,

Decision rule: reject the null hypothesis if $P_{value} <$ level of significance (α)

If the calculated P_{value} is less than the chosen level of significance, we reject the null hypothesis and conclude that the time series data exhibits stationarity. Conversely, if the P_{value} exceeds the significance level, we fail to reject the null hypothesis and conclude that the time series data is non-stationary. The ADF test offers several advantages compared to other stationarity tests. Firstly, it allows for the inclusion of multiple lags of differencing, enabling the detection of more intricate patterns within the data. Secondly, it accommodates a trend term, enabling the consideration of long-term trends in the data.

3.2.3 Trend analysis

Theil-Sen estimator

In many time series literature, researchers often apply the Mann-Kendall test to analyse trends and determine if there is a consistent upward or downward pattern over time. This test assesses the presence of a monotonic trend, indicating a continuous increase or decrease. However, in this research study, we employ a different approach known as the Theil-Sen estimator. This technique, developed by Sen (1968) and Theil (1950) respectively, is a median-based slope estimation method. It offers several advantages over the traditional linear regression as it is non-parametric and more robust. Commonly referred to as the Theil-Sen slope technique or Sen's approach, it provides a reliable means of estimating trends in time series data. Theil-Sen slope is given by the formula:

$$\beta = Median(\frac{x_j - x_i}{j - i}), 1 < i < n, \tag{3.1}$$

where a positive β value, indicates the existence of a rising trend, while a negative β value signifies a declining trend.

3.2.4 Fundamental distributions

In this section, we explore the mathematical foundation of the parent distributions that have been explored in this study and their significance to our study by examining the fundamental principles behind them.

Normal distribution

The normal distribution, often called the Gaussian distribution, is a continuous probability distribution (Marsaglia, 2004), associated with central limit theorem and is given by:

$$N(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{3.2}$$

where μ represent the central value around which the distribution is symmetric and σ measures the spread or dispersion of the distribution.

Gumbel distribution

The probability density function (PDF) of the Gumbel distribution is given by the formula (Coles et al., 2001):

$$G(x) = \frac{1}{\beta} e^{-\left(\frac{x-\mu}{\beta}\right) - e^{-\left(\frac{x-\mu}{\beta}\right)}}, \tag{3.3}$$

where β is a scale parameter which controls the spread or width of the distribution.

Fréchet distribution

The PDF of the Fréchet distribution is given by the formula (Singh et al., 1990):

$$F(x) = \frac{\alpha}{\beta} \left(\frac{x-\mu}{\beta}\right)^{-1-\alpha} e^{-\left(\frac{x-\mu}{\beta}\right)^{-\alpha}}, \tag{3.4}$$

where α is a shape parameter which controls the shape of the distribution.

Weibull distribution

The PDF of the Weibull distribution is given by the formula (Bowden et al., 1983):

$$W(x) = \frac{\alpha}{\beta} \left(\frac{x}{\beta}\right)^{\alpha-1} e^{-(x/\beta)^{\alpha}}. \tag{3.5}$$

3.2.5 Feature scaling

In several machine learning techniques, feature scaling is a significant step in the preparation of the data. Min-Max normalisation is a method that is frequently employed for feature scaling. This approach scales all observations to a particular range, typically between [0,1], based on the minimum and maximum values of the variable.

Min-Max normalisation

Min-Max normalisation is a valuable technique for scaling features to a consistent range, promoting fair and unbiased analysis in machine learning algorithms. By ensuring that all features contribute equally and avoiding dominance based on original ranges, feature scaling enhances the accuracy and efficiency of the models.
Suppose we have n observations in wind speed feature, denoted by an array:

$$\begin{bmatrix} x_1 \\ x_2 \\ . \\ x_n \end{bmatrix} \tag{3.6}$$

Let max represent the highest value in $(x_1, x_2, , ..., x_n)$ and min represent the lowest value in $(x_1, x_2, , ..., x_n)$, then the Min-Max normalisation is given by:

$$f(x) = \frac{x - min}{max - min} \tag{3.7}$$

Original values can be recovered using the function: $f^{-1}(x) = (max - min)x + min$

3.3 Vanilla LSTM network

The Vanilla LSTM network is a form of recurrent neural network that was introduced by Hochreiter and Schmidhuber (1997) to solve the difficulty of processing long-term memory in traditional RNN models. This neural network has numerous gates that enable it to recall information for a long period, in contrast to the traditional RNN models that lose significant previous information.

The capacity of the Vanilla LSTM network to comprehend long-term dependencies is greatly enhanced by the forget gate, which allows it to selectively store or discard information. The Vanilla LSTM network has been demonstrated to give superior memory than many machine learning models, making it particularly suitable for time series forecasting. In order to define the Vanilla LSTM network we let $(x_1, x_2, x_3, x_4..., x_v)$ be input values and $(y_1, y_2, y_3, y_4..., y_v)$ be output values of the existing historical data to be forecasted, then we fit the input values to the five fundamental steps of the LSTM network in a unit, i.e.

$$f_t = g(W_t.[x_v, y_v - 1 + b_f]), \tag{3.8}$$

$$i_t = g(W_i.[x_v, y_v - 1 + b_i]), \tag{3.9}$$

$$O_t = g(W_o.[x_v, y_v - 1 + b_o]), \tag{3.10}$$

$$C_t = f_t c_{t-1} + i_t \tanh(g(W_o.[x_v, y_v - 1 + b_c])), \tag{3.11}$$

$$h_t = O_t \tanh(c_t). \tag{3.12}$$

In equations (3.8) to (3.12), x_v is the vector of the input values, y_v is the vector of the output values, $tanh()$ is the hyperbolic tangent function, b is the bias term and W are weights, C_t is a cell state vector, f_t and O_t are gates of the network.

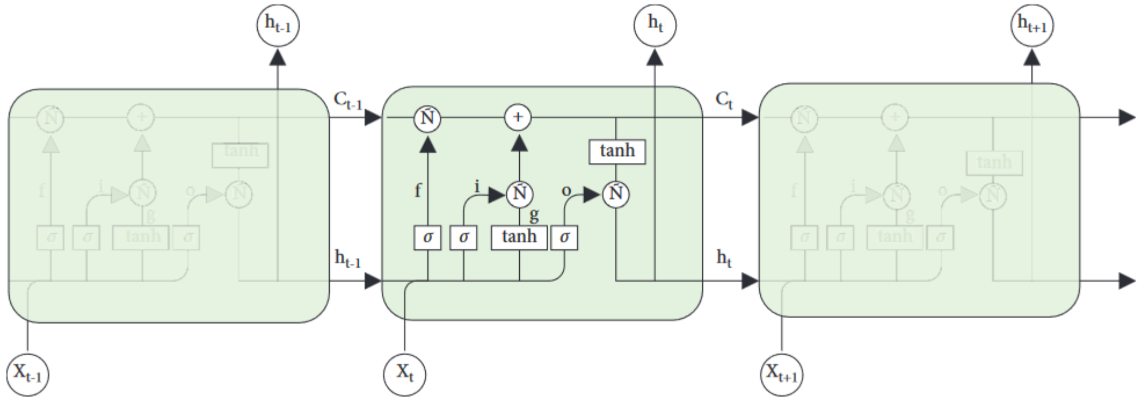


Figure 3.2: The basic structure of the Vanilla LSTM network (Yu et al., 2019).

3.4 Extreme Value Theory model

3.4.1 Generalised Extreme Value Contribution (GEVD)

The GEVD model is an approach for analysing extreme value behaviours in statistics. It focuses on the behaviour of $B_m = Max(X_1, ..., X_m)$, which represents the maximum value in a sequence of independent random variables $(X_1, X_2, ..., X_m)$ with a common distribution $F(x)$. This model is essential for understanding the statistical behaviour of extreme values, and it can be used to make informed decisions based on data analysis (Britten, 2022).

The distribution of B_m can be calculated precisely for all the values of m:

$$\begin{aligned} P_r(B_m \leq z) &= P_r(X_1 \leq z, \ldots, X_m \leq z) \\ &= P_r(X_1 \leq z) \times \ldots \times P_r(X_m \leq z) \\ &= (F(z))^m . \end{aligned} \tag{3.13}$$

The behaviour of $F(z)^m$ as m approaches infinity is examined. However, this analysis is insufficient, as the distribution of B_m will degenerate to a point mass on the upper limit of $F(z)$ for each value of z less than z_{min}. To resolve this issue, a linear re-normalisation of the variable B_m is introduced.

$$B_m^* = \frac{B_m - b_m}{a_m}. \tag{3.14}$$

The Gumbel, Fréchet and Weibull distributions are three popular distribution functions in EVT. Each of these distributions has location and scale parameters, while Fréchet and Weibull have an additional shape parameter. Combining these three distributions creates a family of distribution functions with a probability density function (pdf) form:

$$G(z) = \begin{cases} \exp\left[-\left(1+\xi\left(\frac{z-\mu}{\sigma}\right)\right)^{-\frac{1}{\xi}}\right] & \text{if } \xi \neq 0 \\ \exp\left[-\exp\left(\frac{z-\mu}{\sigma}\right)\right] & \text{if } \xi = 0, \end{cases} \tag{3.15}$$

based on the set $z : 1 + \frac{\varepsilon(z-\mu)}{\sigma} > 0$, with the parameters $-\infty < \varepsilon < \infty$, $-\infty < \mu < \infty$, and $\sigma > 0$. This distribution family is referred to as the GEVD.

3.4.2 Parameter estimation

In this research study, we investigate the estimation of GEVD parameters, specifically focusing on parameter estimation through the MLE (Coles et al., 2001) for $\xi \neq 0$ the likelihood function is:

$$L(\mu, \sigma, \xi) = \prod_{i=1}^{m} \left(\exp\, - \left[1 + \xi \left(\frac{z_{i,r_i} - \mu}{\sigma} \right) \right]^{\frac{-1}{\xi}} \prod_{k=1}^{r_i} \left[1 + \xi \left(\frac{z_{i,k} - \mu}{\sigma} \right) \right]^{\frac{-1}{\xi - 1}} \right) \tag{3.16}$$

and for $\xi = 0$

$$L(\mu, \sigma, \xi) = \prod_{i=1}^{m} \left(\exp \left\{ - \exp \left[- \frac{z_{i,r_i} - \mu}{\sigma} \right] \right\} \prod_{k=1}^{r_i} \left[\sigma^{-1} \exp \left(- \frac{z_{i,k} - \mu}{\sigma} \right) \right] \right). \tag{3.17}$$

In the context of this study, the optimisation of log-likelihoods yields parameter estimates. These estimated parameters align with those characterising the GEVD for block maxima but encompass a broader range of extreme observations.

3.4.3 Return levels

For the GEVD in equation (3.15), we can determine the return level, denoted as Z_p, associated with a specific return period for each set of p observations (Beirlant et al., 2006). This relationship can be expressed through the following formula:

$$Z_p = \begin{cases} \mu - \frac{\sigma}{\xi} \left(1 - (-\log(1-p))^{-\xi} \right) & \text{if } \xi \neq 0 \\ \mu - \sigma \log(-\log(1-p)) & \text{if } \xi = 0 \end{cases} \tag{3.18}$$

3.4.4 Evaluations metrics for the Vanilla LSTM

In evaluating the accuracy of the Vanilla LSTM network, this study employs various metrics to measure this precision. The selected evaluation measures include the well-established Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE). Each metric is pivotal in unveiling distinct aspects of Vanilla LSTM accuracy. The MAE provides insights into the average magnitude of errors between predicted and actual values, while the RMSE underscores precision by capturing the square root of the mean of squared errors. Additionally, the MAPE offers a percentage-wise assessment of the average difference between predicted and actual values.

The mathematical expressions for these metrics are as follows:

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n} |y_t - \hat{y}_t|, \tag{3.19}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (y_t - \hat{y}_t)^2}{n}}, \tag{3.20}$$

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n} \left|\frac{y_t - \hat{y}_t}{y_t}\right| \times 100\%, \tag{3.21}$$

where y_t signifies the actual values at time t, $\hat{y}_t$ represents the predicted model values, $\bar{y}$ denotes the mean of the actual observations, and n is the sample size of the total observations. These metrics offer comprehensive insights into the accuracy and reliability of our predictive models (Naser and Alavi, 2020; Herrera et al., 2010)

3.4.5 Evaluations metrics for the GEVD

A statistical model adequacy for a particular set of data is evaluated using the AIC. This measure was developed by Akaike (1973). The AIC has gained in popularity in areas like EVT, machine learning and econometrics. It has been demonstrated to be successful in preventing overfitting.

The following is a mathematical formula that defines the AIC:

$$\text{AIC} = -2(\log(L)) + 2K, \tag{3.22}$$

where $\log(L)$ is the likelihood of the model and k represents the total count of the model parameters. The second component of the equation indicates that the AIC penalises models with more parameters. A better fit is indicated by lower AIC value.

The BIC and AIC are interrelated, as they serve the common purpose of assessing model performance. The BIC imposes a harsher penalty on models with several parameters than the AIC. According to Kulkarni and Desai (2017), the BIC is a near approximation of the Bayes factor between two models.

The following is a mathematical formula that defines the BIC:

$$\text{BIC} = -2(\log(L)) + K\log(N), \tag{3.23}$$

where the sample size is N. Additionally, the BIC penalises models with more parameters, but more severely than the AIC. As a result, compared to the AIC, the BIC favours more straightforward models.

Chapter 4

Results and discussion

4.1 Introduction

This chapter focuses on the analysis and discussion of the results obtained from this research study. This chapter is structured into distinct sections, including descriptive statistics, stationarity test, Nonparametric trend analysis, machine learning and EVT results.

4.2 Data wrangling

The dataset appeared to be devoid of any missing or null values, with a collecting period spanning from 2016 to 2022. Additionally, feature engineering was conducted to divide the months into distinct seasons, namely Summer, Winter, Autumn and Spring. This approach was undertaken to enhance our comprehension of the trends and patterns prevalent within our designated research area.

4.3 Descriptive statistics

In this study, various descriptive statistics were used to evaluate wind patterns in Polokwane. These statistics include the minimum, maximum, mean, standard deviation, median, kurtosis and skewness.

Table 4.1: Wind speed summary statistics of Polokwane.

	min	median	mean	max	skewness	kurtosis	Std Deviation
Wind Speed	0.20	7.96	8.15	22.86	0.683	4.01	2.858

In the Table 4.1, the minimum wind speed recorded in Polokwane stands at 0.20 metres per second, representing the lowest observed wind speed during the data collection period. Conversely, the maximum wind speed recorded is 22.86 metres per second, indicating the highest observed wind speed. The median, at 7.96 metres per second, which signifies the middle point in the distribution, dividing the data into two halves. The mean wind speed, obtained as 8.15 metres per second, provides an average representation of wind conditions in Polokwane.

Furthermore, the skewness value of 0.683 suggests a positively skewed distribution, indicating occasional high wind speed events in the region. The kurtosis value of 4.01 reveals that the distribution is leptokurtic, meaning it has heavier tails and is more peaked around the mean compared to a normal distribution. Lastly, the standard deviation, which is 2.858 metres per second, quantifies the spread or variability of wind speed values around the mean.

4.3.1 Seasons summary statistics

In our research study, we conducted a detailed analysis of descriptive statistics for each season within our focus area. The results, which are presented in the accompanying Table 4.2, offering valuable insights relevant to energy generation and management, recreation and outdoor activities, emergency preparedness and urban planning.

Table 4.2: Seasons summary statistics.

<table>
<tr><th colspan="7">Summer</th></tr>
<tr><td>min</td><td>median</td><td>mean</td><td>Std Dev</td><td>max</td><td>skewness</td><td>kurtosis</td></tr>
<tr><td>2.38</td><td>7.34</td><td>7.36</td><td>2.20</td><td>13.55</td><td>0.245</td><td>-0.088</td></tr>
<tr><th colspan="7">Winter</th></tr>
<tr><td>min</td><td>median</td><td>mean</td><td>Std Dev</td><td>max</td><td>skewness</td><td>kurtosis</td></tr>
<tr><td>0.2</td><td>7.96</td><td>8.35</td><td>3.45</td><td>22.87</td><td>0.913</td><td>1.321</td></tr>
<tr><th colspan="7">Autumn</th></tr>
<tr><td>min</td><td>median</td><td>mean</td><td>Std Dev</td><td>max</td><td>skewness</td><td>kurtosis</td></tr>
<tr><td>2.06</td><td>8.90</td><td>8.78</td><td>2.91</td><td>18.21</td><td>0.312</td><td>-0.149</td></tr>
<tr><th colspan="7">Spring</th></tr>
<tr><td>min</td><td>median</td><td>mean</td><td>Std Dev</td><td>max</td><td>skewness</td><td>kurtosis</td></tr>
<tr><td>2.38</td><td>7.34</td><td>7.36</td><td>2.20</td><td>13.55</td><td>0.245</td><td>-0.088</td></tr>
</table>

The analysis of seasonal wind speed statistics in Table 4.2 reveals distinctive patterns in each season. Summer exhibits moderate wind conditions, with a median wind speed of 7.34 metres per second, indicating a reasonably efficient period for wind power generation. Winter stands out as the most wind efficient season with a higher median wind speed of 7.96 metres per second, although it also presents the potential for extreme wind events, as indicated by a positively skewed distribution and higher kurtosis. Autumn showcased a favourable median wind speed of 8.90 metres per second, offering another promising season for wind power generation, while its distribution remains relatively stable. In addition to this, spring and summer display moderate wind speeds, further contributing to a balanced seasonal profile.

4.4 Diagnostic plots

Figure 4.1 illustrates diagnostic plots for our response feature, to check statistical underlying assumptions of the data, including normality and homoscedasticity.

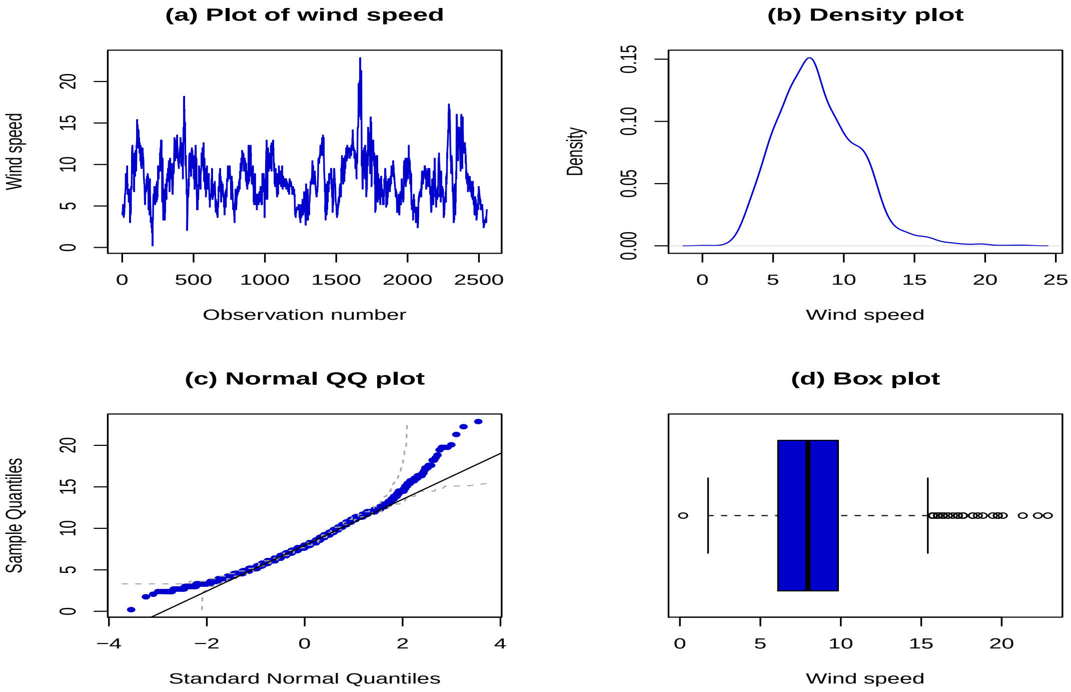


Figure 4.1: Diagnostic plots.

The top left quadrant in Figure 4.1 is a time series plot depicting the wind pattern from 2016 to 2022. Top right quadrant is a density plot displaying that the data is positively skewed. This means that, most of the time, the wind tends to blow at lower average speed, with occasional burst. The bottom left quadrant normal Q-Q plot shows that the upper points depart from a straight line, proving that the data is right-skewed. The bottom right quadrant, the box plot reveals that there are outliers, indicating few low wind occurrence events in Polokwane.

4.5 Test for stationarity

The ADF test is deployed to determine whether the wind speed in Polokwane is stationary or non-stationary. The ADF test results are as follows:

Step 1: Hypothesis

H_0: Time series data exhibit a unit root or stochastic trend, indicating it is non-stationary.

H_1: Time series data is stationary.

Step 2: Level of significance

$\alpha = 5\% = 0.05$

Step 3: Computation of p-value

$P_{value} \leq 0.01$.

Step 4: Decision

Since $P_{value} < \alpha$ we reject the null hypothesis and conclude that the data does not exhibit a unit root or stochastic trend, meaning that the data is stationary.

4.6 Nonparametric trend estimation

Table 4.3: Theil-Sen estimator results.

Estimator	**Value**
Theil-Sen estimator	-0.0002
Test Statistics	-3.4644
Degree of freedom	2555

Table 4.3 presents the outcomes of the Theil-Sen estimator within the framework of our wind speed research study. The Theil-Sen estimator value of -0.0002 indicates the potential linear relationship between the two variables, wind speed and date time. The negative test statistic of -3.4644 suggests that

the estimated slope differs from zero, indicating a potential negative association between the two variables. With a degree of freedom $(n-2) = 2555$, this statistical analysis captures the variability in the data. In the context of wind power generation, this slightly negative slope could imply a slight reduction in wind power output or efficiency as wind speed increases, though the magnitude of this effect is minimal (-0.0002).

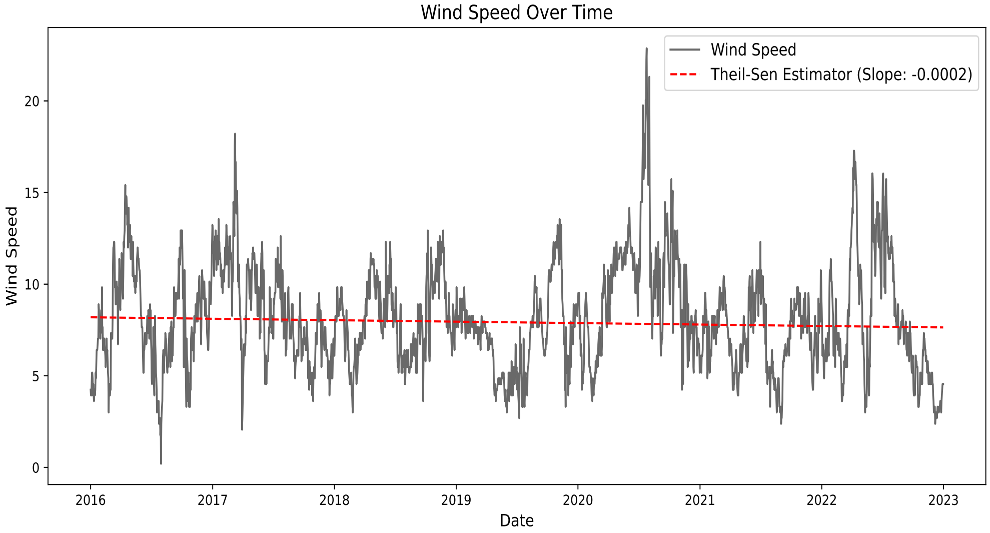


Figure 4.2: Trend slope plot.

Consequently, this minimal Theil-Sen value estimate from Table 4.3 implies a lower frequency of maintenance and repair requirements for wind turbines, aligning with the stability and reliability of wind energy generation systems. This relationship described is supported by the fitted Theil-Sen red dotted slope depicted in Figure 4.2, where a subtle downward slope is observed. Notably, this unclear relationship may not be easily identifiable by visual inspection alone, emphasising the importance of employing the slope estimator as a critical tool in uncovering and quantifying this relationship within the data.

4.7 Machine learning results

4.7.1 Model parameter setting

Table 4.4: Parameter setting for LSTM network.

Parameters	**Values**
Number of units	4
Activation function	$tanh$
Optimizer	Adam
Epochs	200
Splitting ratio	7:3

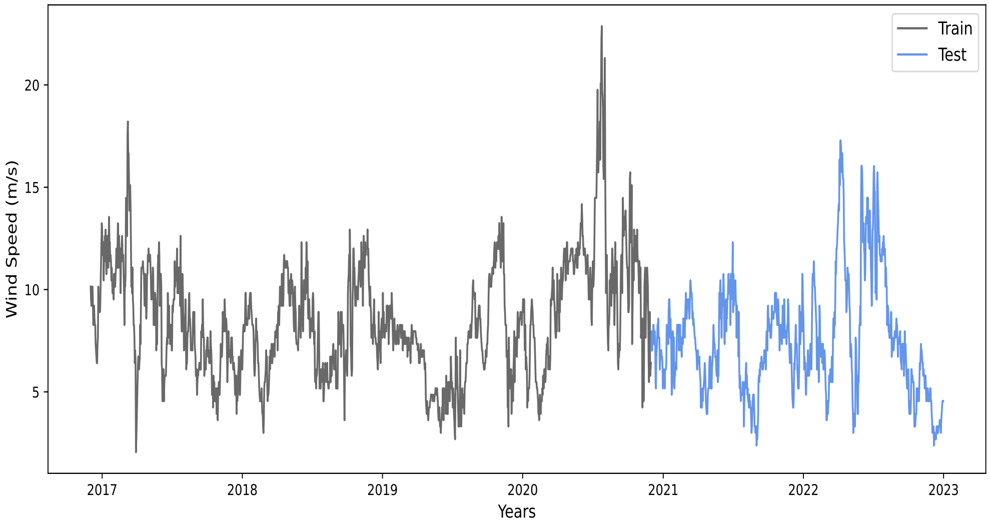


Figure 4.3: Splitting data.

Table 4.4, presents the key parameter settings employed in the configuration of the Vanilla LSTM network, a critical component in our research findings. These parameters play a pivotal role in the model performance and effectiveness in capturing intricate temporal patterns within our dataset. Table 4.4 indicates how the Vanilla LSTM network was structured with a relatively modest number of units, specifically four. This choice reflects a balance between model complexity and computational efficiency, allowing for the extraction of meaningful features from the data while avoiding overfitting. The activation function employed was the hyperbolic tangent $(tanh)$, which facilitates the model capability to capture both positive and negative correlations in the wind speed data. We employed the Adam optimizer, a widely adopted optimisation algorithm, to fine-tune the model weights during training. Furthermore, the training process was carried out over 200 epochs, indicating the number of iterations through the dataset to optimise the model performance.

Finally, the data was divided into a training set and a testing set using a splitting ratio of 7:3, ensuring that the model had sufficient data for learning while maintaining a robust validation process to assess its generalisation performance. In Figure 4.3, the grey line represents 70% of the training data, while the blue line represents 30% of the testing observations. This indicates a division between the data used for training and the data used for testing. This separation is important in machine learning and statistical analysis to evaluate the performance and accuracy of a model, facilitating the assessment of how well the model generalises to unseen data. These parameter settings are instrumental in achieving reliable and informative results within our research of study.

4.7.2 Vanilla LSTM network

The Vanilla LSTM network stands out as a powerful algorithm for wind speed, rainfall and temperature prediction, boasting wide-ranging applications from renewable energy administration to environmental development planning. As we reveal the results stemming from our research study, it is essential to emphasise that these forecasts extend beyond mere numerical data points. They possess the capacity to instigate significant shifts in decision-making procedures across diverse industries. The results are presented in Figure 4.4.

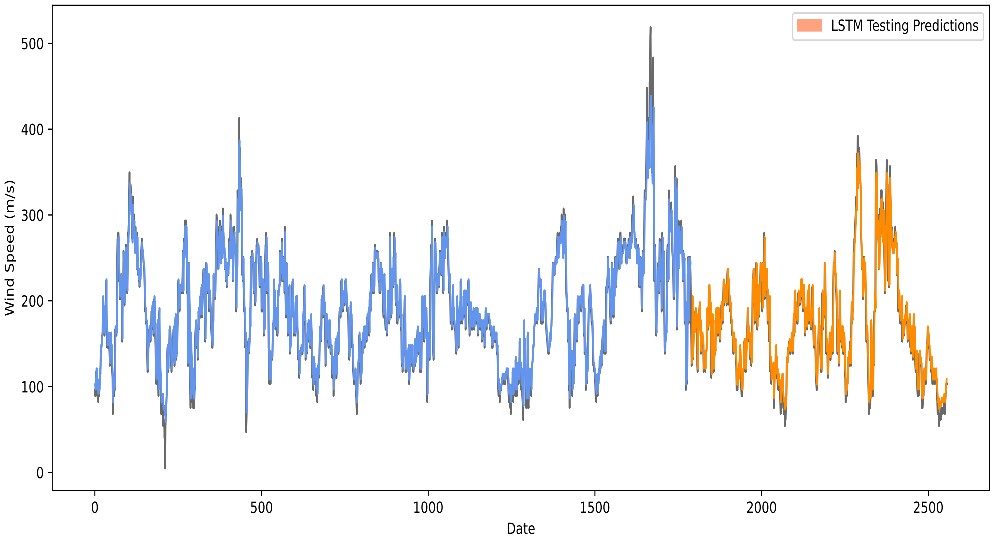


Figure 4.4: LSTM network predictions.

Figure 4.4 presents the Vanilla LSTM network predictions for the training and testing sets, illustrating the proximity of the predicted values to the actual values. The blue line represents the model predictions for the training set, while the orange line represents the predictions for the testing set. This plot illustrates the model capacity to generalise effectively to dataset with similar characteristics. These predictions highlight the model proficiency in providing reliable wind speed forecasts, a significant objective within our research study.

4.7.3 Evaluation metrics

Table 4.5: Vanilla LSTM evaluation metrics.

Training Evaluation metrics				
	MAE	RMSE	MAPE	Accuracy
LSTM network	0.168	0.235	0.101	0.86
Testing Evaluation metrics				
	MAE	RMSE	MAPE	Accuracy
LSTM network	0.147	0.207	0.090	0.89

Table 4.5 displays a comprehensive assessment of the performance of the Vanilla LSTM network through various evaluation metrics for both training and testing phases. In the training phase, we observe that the MAE stands at 0.168, indicating the average magnitude of prediction errors. The RMSE, a measure of the model precision, is 0.235. Additionally, the MAPE reflects a low value of 0.101, underscoring the model accuracy in forecasting. The overall accuracy of the Vanilla LSTM network on the testing data is recorded at 89% with an increase of 3% from the training data, denoting a high level of correctness in predictions. This robust performance reinforces confidence in the model reliability for real-world scenarios.

The MAE for testing is 0.147, indicative of the model ability to guarantee good generalization to unseen data. Similarly, the RMSE remains low at 0.207, emphasising precision in forecasting. Moreover, the MAPE margin decreased to 0.090, representing a good level of accuracy. It is crucial to note that, when dealing with sequential data or time series, accuracy alone may not be sufficient. Metrics such as MAE, RMSE and MAPE are more important to consider for regression tasks. These evaluation metrics are essential to demonstrate the robustness and effectiveness of the Vanilla LSTM network in wind speed forecasting, affirming its suitability for real-world applications.

4.8 Extreme value analysis results

4.8.1 modelling GEVD

Table 4.6: GEVD parameter estimation.

m	$\hat{\mu}$	$SE(\hat{\mu})$	$\hat{\sigma}$	$SE(\hat{\sigma})$	$\hat{\xi}$	$SE(\hat{\xi})$	95% CI ($\hat{\xi}$)
1	8.910	0.382	1.387	3.964	-0.453	0.633	(-0.374,0.129)
2	9.285	0.219	2.566	0.154	-0.037	0.048	(-0.131,0.057)
3	9.152	0.176	2.526	0.124	-0.045	0.039	(-0.122,0.031)
4	9.025	0.151	2.509	0.107	-0.055	0.034	(-0.121,0.011)
5	8.918	0.134	2.496	0.094	-0.062	0.029	(-0.120,-0.004)

Table 4.6 presents the parameter estimation results for a GEVD model for different block sizes represented by m values, ranging from 1 to 5. The results unveiled valuable insights into the statistical characteristics of extreme events, which are of significant importance for understanding and modelling rare and extreme occurrences. The estimated location parameter $\hat{\mu}$ exhibited a decreasing trend as m increased, indicating a shift towards lower central tendencies in the extreme event data. This suggests that larger blocks tend to have extreme events with lower average values. Moreover, the standard errors $SE(\hat{\mu})$, associated with $\hat{\mu}$ decreased as m increased, reflecting a higher level of confidence in the location parameter estimate for larger block sizes.

The scale parameter $\hat{\sigma}$ remained relatively stable across different block sizes, suggesting that the spread or variability of extreme events did not vary significantly with block sizes. The shape parameter $\hat{\xi}$ plays a significant role, which dictates the tail behaviour of the distribution. As m increased, $\hat{\xi}$ became more negative, indicating a transition towards a heavier-tailed distribution for larger blocks. These findings contribute valuable insights into characterising extreme events and tail behaviour, aiding in selecting the optimal block size for extreme event prediction and risk assessment.

4.8.2 Goodness of fit GEVD

Table 4.7: Goodness of fit evaluation metrics.

m	$-\log(\text{Likelihood})$	AIC	BIC
1	1184.656	2375.312	2382.605
2	**419.271**	**844.542**	**853.914**
3	623.655	1253.311	1263.899
4	827.407	1660.813	1672.264
5	1030.059	2066.118	2078.239

Table 4.7 presents a comprehensive evaluation of the goodness of fit metrics for different block sizes. The metrics include the negative log-likelihood, AIC and BIC. These metrics are crucial in assessing how well each block size can represent the original data while considering the trade-off between model complexity and goodness of fit. In this assessment, we aim to identify the most suitable block size to represent our data.

The results in Table 4.7 reveal that as m increases, both the negative log-likelihood and the associated AIC and BIC values increases from block size two to five. This trend indicates that larger values for block size result in progressively poorer fits to the observed data. However, the AIC and BIC metrics take into account model complexity, penalising models with more parameters. While higher m values may capture more complexity, they may also introduce overfitting issues. In the quest for the best model, it is essential to strike a good balance between model goodness of fit and complexity.

In this context, $GEVD_{m=2}$ stands out as the best fit as it offers a substantially lower AIC and BIC compared to other models, while maintaining a reasonably low negative log-likelihood. Thus $m = 2$ block strikes an effective balance between explaining the data and avoiding excessive complexity, making it the most favourable choice among the options presented.

4.9 $GEVD_{m=2}$ results

4.9.1 Diagnostic plots for GEVD

One other method for determining the best block size involves a comprehensive approach that combines visual inspection of the data using graphical diagnostic plots with the insights derived from the results presented in Table 4.7, particularly emphasising the model with the smallest AIC and BIC values.

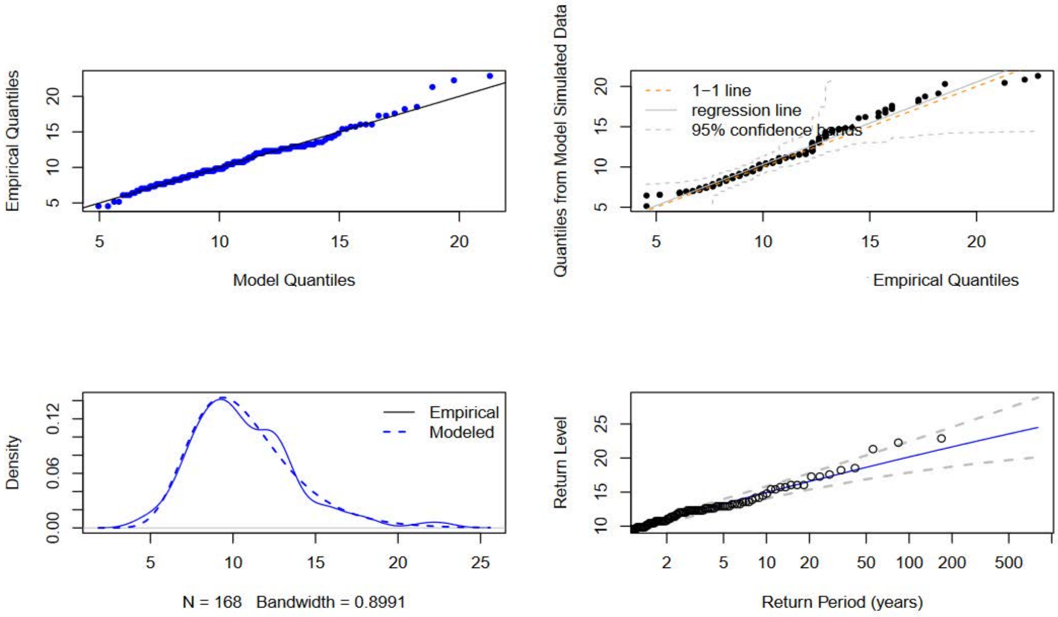


Figure 4.5: $GEVD_{m=2}$ diagnostic plots.

Figure 4.5 offers compelling evidence for the $GEVD_{m=2}$ suitability for describing wind speed behaviour. This is evident through strong alignment in the P-P and Q-Q plots, a clear increasing trend in the return level plot, and a close match between the fitted and empirical density in the density plot. This robust agreement between the model and empirical data underscores the model reliability. Furthermore, it underscores its potential to enhance our understanding in prediction of extreme wind events, a critical factor in wind power generation.

4.9.2 Return levels for GEVD

Table 4.8: Return levels.

5 years	20 years	50 years	100 years	200 years	250 years	300 years
13.030	16.505	18.612	20.144	21.631	22.102	22.893

Table 4.8 provides essential information regarding return levels for different return periods, ranging from 5 years to 300 years. These return levels represent critical values within a distribution, delineating the expected magnitude of extreme events over specified time intervals.

The return level for a 5-year period is determined to be 13.030 metres per second. This implies that, on average, an event of this magnitude is anticipated to occur approximately once every 5 years. As we extend the return period to 20 years, the return level increased to 16.505 metres per second, indicating that such an event is expected less frequently within a 20-years time period.

The 250-year return level denotes an exceedingly uncommon occurrence, anticipated to transpire approximately once every 250 years and it has been found to be 22.102 metres. Going beyond this rarity, the 300 year return level signifies an event of even greater infrequency, projected to manifest approximately once every three centuries. In light of the maximum recorded wind speed of 22.86 meters per second, as detailed in Table 4.1, the 300 year return level serves as a pivotal benchmark for evaluating the potential risk associated with the most exceptional wind events that may be encountered in the Polokwane region. Notably, this event is expected to occur during the winter season, as indicated by the findings in Table 4.2, which presents seasonal summary statistics.

4.10 Summary of the chapter

In conclusion, this chapter effectively addressed the objectives of this research study by applying the Vanilla LSTM algorithm and GEVD to model wind speed data. The successful application of these methodologies demonstrated their adaptability to the complexities of the dataset. The assessment of the Vanilla LSTM predictive performance revealed its efficiency in short-term wind speed forecasting, providing valuable insights for real-time applications. The GEVD analysis affirmed its suitability for long-term modelling of extreme wind speeds, contributing a robust framework for understanding extreme events.

The determination of the optimal block size for modelling extreme events using GEVD emerged as a critical contribution, offering practical guidance for future applications. The optimal block size was selected based on low AIC and BIC value. Additionally, $GEVD_{m=2}$ was employed to estimate return levels of extreme wind speeds and corresponding return periods, providing essential information for risk assessment and infrastructure planning. Combining findings from both methodologies, this study uncovered overarching trends, validating the effectiveness of the Vanilla LSTM in short-term forecasting and emphasising the $GEVD_{m=2}$ robustness in capturing long-term extreme events. These findings stand out from existing literature as we have employed both the Vanilla LSTM and GEVD within a single study, enabling a comprehensive examination of both short and long-term predictions for robust data modelling. This enhances and enriches our research study, making it stand out in the field of forecasting and environmental science.

Chapter 5

Conclusion

5.1 Introduction

This chapter covers the conclusion and recommendations of this research study. It is divided into four subsections, namely, conclusion, recommendations, limitations and future studies.

5.2 Conclusion

This subsection provides a comprehensive and in-depth analysis of wind speed forecasting in Polokwane, shedding light on the various statistical properties and seasonal variations within the region. The findings revealed a wide range of wind speeds, with the lowest recorded speed at 0.20 metres per second and the highest at 22.86 metres per second. The descriptive statistics, skewness, kurtosis and standard deviation, revealed the presence of occasional high wind speed events and the distribution was found to be of leptokurtic nature.

Seasonal analysis has highlighted the efficiency of wind power generation during certain periods, such as winter and autumn, while also indicating the potential for extreme wind events. Moreover, the examination of time series data has shown that it is stationary using ADF test, a crucial aspect for accurate modelling and prediction.

The evaluation of $GEVD_{m=2}$, identified as the optimal EVT model in comparison to $GEVD_{m=1,3,4,5}$, which provided valuable insights into the statistical characteristics of extreme events. This is especially notable in terms of the location, scale and shape parameters. Furthermore, the Vanilla LSTM network performance has demonstrated its ability to provide accurate wind speed forecasts with low MAE, RMSE, MAPE and a predictive accuracy of 89% for testing data , offering promising prospects for practical applications.

The assessment of a $GEVD_{m=2}$, which was found to be the best EVT model has contributed significant insights into the statistical characteristics of extreme events, particularly regarding to location, scale and shape parameters. These findings are invaluable for developing accurate models for extreme event prediction and risk assessment. In addition, the determination of return levels for return periods of 5, 15, 25, 50, and 100 years has revealed a progressive escalation with longer return periods. This information is of paramount importance for conducting risk assessments and facilitating decision-making processes, allowing stakeholders to assess both the relative infrequency and magnitude of extreme wind events over a designated time intervals.

Overall, these research study findings offers a solid foundation for wind power generation. Ultimately, they serve as a valuable insight for stakeholders in the renewable energy sector and environmental planning, contributing to more sustainable and informed decision-making in Polokwane.

5.3 Recommendations

The research study offers a set of recommendations in the domain of energy management and sustainability. The following recommendations were made:

1. Employ advanced machine learning techniques, such as Convolutional Neural Networks (CNN). This method application in wind forecasting is considered beneficial due to its capability to automatically extract hierarchical features from geographical data. Unlike traditional methods, CNN can capture complex relationships within meteorological variables and geographical features, thereby enhancing the accuracy and efficiency of wind forecasting.

2. Investigate modern EVT techniques, such as bGEVD, present a refined approach to analysing extreme events in wind speed data. The application of bGEVD allows for a more clear examination of extreme events, providing insights into tail behaviour and facilitating a better understanding of rare, high-impact occurrences. This method can extends the toolkit for extreme event analysis, enabling researchers to capture and interpret extreme wind events more accurately. By incorporating bGEVD, future research can gain a deeper understanding of the tail behaviour of wind speed distributions, thereby improving the assessment of potential risks associated with extreme wind events.

3. Expand the geographical scope to encompass all nine provinces of South Africa for a clear understanding of regional wind patterns. Each province has unique geographical features, climate conditions, and wind characteristics that significantly influence wind power potential. A broader geographical scope will enable the research to capture diverse wind patterns, ensuring a more robust and representative analysis.

4. Encourage collaboration and knowledge sharing among research institutions, government entities and industry stakeholders. By encouraging collaboration among research institutions, government entities and industry stakeholders, the research outcomes and data can be leveraged collectively. The exchange of knowledge and data facilitates can offer a more comprehensive understanding of the challenges and opportunities in wind power generation.

5.4 Limitations of the study

Wind is a magnificent, boundless and cost-free clean source of renewable energy for electricity generation. In light of this, it is essential to recognise specific constraints within wind power generation. This research study is conducted in the capital city of Limpopo, which is a semi-developed urban area. This geographical focus presents a limitation, as addressing the issue of load shedding requires a broader perspective encompassing all cities across South Africa.

Moreover, there is scarcity of comprehensive literature globally, specifically within South Africa, that integrates both machine learning and Extreme Value Theory (EVT) in a single research study for the purpose of wind speed forecasting. This lack of literature in this specific domain poses a challenge in terms of comparing and contextualising the findings of this research study. However, these limitations emphasise the urgency of further research and collaboration to tackle the multifaceted challenges associated with wind power generation and load shedding mitigation in a more holistic manner.

5.5 Future studies

For future investigations, we recommend that fellow researchers replicate this experiment employing modern EVT techniques and machine learning. Incorporate techniques such as bGEVD or the time-varying threshold Generalized Pareto Distribution (GPD) methodologies. Researchers should also apply a variety of machine learning techniques and deep learning algorithms within the field of artificial intelligence to develop a robust algorithm suitable for their specific area of study.

Additionally, it is advisable to redirect other researchers to focus on all the provinces of South Africa namely, Western Cape, Eastern Cape, Northern Cape, Free State, KwaZulu-Natal, Gauteng, North West, Mpumalanga, and Limpopo. This strategy is aimed at achieving an accurate equilibrium between electricity demand and supply, thereby fostering a future marked by economic prosperity and environmental sustainability.

References

Abdulali, B. A. A., Abu Bakar, M. A., Ibrahim, K., Mohd Ariff, N., et al. (2022). Extreme value distributions: An overview of estimation and simulation. *Journal of Probability and Statistics*, **2022**.

Agilan, V. and Umamahesh, N. (2017). What are the best covariates for developing non-stationary rainfall intensity-duration-frequency relationship? *Advances in Water Resources*, **101**, 11–22.

Ahmed, K. M., Khan, M. A., Siddiqui, I., Khan, S., Shoaib, M., and Zia, I. (2022). Wind speed prediction from site meteorological data using artificial neural network. *Institute of Electrical and Electronics Engineers*, 1–8.

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle–in: Second international symposium on information theory. *Budapest: Academiai Kiado*.

Al Dhaheri, K., Woon, W. L., and Aung, Z. (2017). Wind speed forecasting using statistical and machine learning methods: A case study in the uae. *Springer*, 107–120.

Al-Ghussain, L. (2019). Global warming: review on driving forces and mitigation. *Environmental Progress & Sustainable Energy*, **38** (1), 13–21.

Antonini, E. G. and Caldeira, K. (2021). Atmospheric pressure gradients and coriolis forces provide geophysical limits to power density of large wind farms. *Applied Energy*, **281**, 116048.

ARENT, D. J., WISE, A., AND GELMAN, R. (2011). The status and prospects of renewable energy for combating global warming. *Energy Economics*, **33** (4), 584–593.

ARORA, P., KUMAR, H., AND PANIGRAHI, B. (2018). A comparative study for short term wind speed forecasting using statistical and machine learning approaches. *2018 2nd IEEE International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES)*, 200–205.

AWEA (2017). Awea u.s. wind industry annual market report: Year ending 2016. washington, d.c. *American Wind Energy Association*.

BANDERKER, S. (2022). The perceived psychosocial and economic impact of load-shedding on employees in selected small micro medium enterprises. *Masters dissertation, University of the Western Cape*.

BEIRLANT, J., GOEGEBEUR, Y., SEGERS, J., AND TEUGELS, J. L. (2006). Statistics of extremes: theory and applications. *John Wiley & Sons*.

BHAGWANDIN, L. (2017). Multivariate extreme value theory with an application to climate data in the western cape province. *Masters dissertation, University of Cape Town*.

BHASKARAN, S., VERMA, A. S., GOUPEE, A. J., BHATTACHARYA, S., NEJAD, A. R., AND SHI, W. (2023). Comparison of extreme wind and waves using different statistical methods in 40 offshore wind energy lease areas worldwide. *Energies*, **16** (19), 6935.

BOKDE, N., FEIJÓO, A., VILLANUEVA, D., AND KULAT, K. (2019). A review on hybrid empirical mode decomposition models for wind speed and wind power prediction. *Energies*, **12** (2), 254.

BOTZEN, W. W., DESCHENES, O., AND SANDERS, M. (2019). The economic impacts of natural disasters: A review of models and empirical studies. *Review of Environmental Economics and Policy*.

BOWDEN, G., BARKER, P., SHESTOPAL, V., AND TWIDELL, J. (1983). The weibull distribution function and wind power statistics. *Wind Engineering*, 85–98.

BRABSON, B. AND PALUTIKOF, J. (2000). Tests of the generalized pareto distribution for predicting extreme wind speeds. *Journal of Applied Meteorology and Climatology*, **39** (9), 1627–1640.

BRITTEN, G. L. (2022). Extreme value distributions describe interannual variability in the seasonal north atlantic phytoplankton bloom. *Limnology and Oceanography Letters*, **7** (3), 269–276.

CARLEY, S., ENGLE, C., AND KONISKY, D. M. (2021). An analysis of energy justice programs across the united states. *Energy Policy*, **152**, 112219.

CASTILLO, E., HADI, A., BALAKRISHNAN, N., AND SARABIA, J. (2005). Extreme value and related models in engineering and science applications. *John Wiley & Sons, Hoboken, New Jersey*, **179** (180), 31.

CASTILLO-MATEO, J., ASÍN, J., CEBRIÁN, A. C., MATEO-LÁZARO, J., AND ABAURREA, J. (2023). Bayesian variable selection in generalized extreme value regression: modeling annual maximum temperature. *Mathematics*, **11** (3), 759.

CHAPMAN, S., BACON, J., BIRCH, C. E., POPE, E., MARSHAM, J. H., MSEMO, H., NKONDE, E., SINACHIKUPO, K., AND VANYA, C. (2023). Climate change impacts on extreme rainfall in eastern africa in a convection-permitting climate model. *Journal of Climate*, **36** (1), 93–109.

COHEN, A. J., BRAUER, M., BURNETT, R., ANDERSON, H. R., FROSTAD, J., ESTEP, K., BALAKRISHNAN, K., BRUNEKREEF, B., DANDONA, L., DAN-

DONA, R., ET AL. (2017). Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the global burden of diseases study 2015. *The lancet*, **389** (10082), 1907–1918.

COLES, S., BAWA, J., TRENNER, L., AND DORAZIO, P. (2001). An introduction to statistical modeling of extreme values. *London: Springer*.

CSIR (2023). Annual statistics on power generation in south africa for 2022 (1 january 2022 to 31 december 2022), including loadshedding and energy availability factor data. *South Africa: CSIR*.

DANIEL, L. O., SIGAUKE, C., CHIBAYA, C., AND MBUVHA, R. (2020). Short-term wind speed forecasting using statistical and machine learning methods. *Algorithms*, **13** (6), 132.

DENG, X., ZHANG, F., WANG, L., RUI, W., LONG, F., ZHAO, Y., CHEN, D., AND DING, W. (2014). Airborne fine particulate matter induces multiple cell death pathways in human lung epithelial cells. *Apoptosis*, **19**, 1099–1112.

DICKEY, D. A. AND FULLER, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, **74** (366a), 427–431.

DIFFENBAUGH, N. S. AND BURKE, M. (2019). Global warming has increased global economic inequality. *Proceedings of the National Academy of Sciences*, **116** (20), 9808–9813.

DIRIBA, T. A., DEBUSHO, L. K., AND BOTAI, J. (2015). Modeling extreme daily temperature using generalized pareto distribution at port elizabeth, south africa. *Sabinet African Journals*, **2015** (1), 41–48.

DONAGHY, T. Q., HEALY, N., JIANG, C. Y., AND BATTLE, C. P. (2023). Fossil fuel racism in the united states: How phasing out coal, oil, and gas can protect communities. *Energy Research & Social Science*, **100**, 103104.

DOTTORI, F., SZEWCZYK, W., CISCAR, J.-C., ZHAO, F., ALFIERI, L., HIRABAYASHI, Y., BIANCHI, A., MONGELLI, I., FRIELER, K., BETTS, R. A., ET AL. (2018). Increased human and economic losses from river flooding with anthropogenic warming. *Nature Climate Change*, **8** (9), 781–786.

ELSARAITI, M. AND MERABET, A. (2021). A comparative analysis of the arima and lstm predictive models and their effectiveness for predicting wind speed. *Energies*, **14** (20), 6782.

FISCHEREIT, J., BROWN, R., LARSÉN, X. G., BADGER, J., AND HAWKES, G. (2022). Review of mesoscale wind-farm parametrizations and their applications. *Boundary-Layer Meteorology*, **182** (2), 175–224.

GENG, D., ZHANG, H., AND WU, H. (2020). Short-term wind speed prediction based on principal component analysis and lstm. *Applied Sciences*, **10** (13), 4416.

HAINES, A., KOVATS, R. S., CAMPBELL-LENDRUM, D., AND CORVALÁN, C. (2006). Climate change and human health: impacts, vulnerability, and mitigation. *The Lancet*, **367** (9528), 2101–2109.

HERRERA, M., TORGO, L., IZQUIERDO, J., AND PÉREZ-GARCÍA, R. (2010). Predictive models for forecasting hourly urban water demand. *Journal of hydrology*, **387** (1-2), 141–150.

HOCHREITER, S. AND SCHMIDHUBER, J. (1997). Long short-term memory. *Neural Computation*, **9** (8), 1735–1780.

IEA (2020). Global investment in clean energy 2020: Europe. *International Energy Agency*.
URL: *https://www.iea.org/*

JENSEN, J. P. AND SKELTON, K. (2018). Wind turbine blade recycling: Experiences, challenges and possibilities in a circular economy. *Renewable and Sustainable Energy Reviews*, **97**, 165–176.

KIRBAS, I. AND KEREM, A. (2016). Short-term wind speed prediction based on artificial neural network models. *Measurement and Control*, **49** (6), 183–190.

KULKARNI, S. AND DESAI, S. (2017). Classification of gamma-ray burst durations using robust model-comparison techniques. *Astrophysics and Space Science*, **362** (4), 70.

MADIBA, T., BANSAL, R., MBUNGU, N., BETTAYEB, M., NAIDOO, R., AND SITI, M. (2022). Under-frequency load shedding of microgrid systems: a review. *International Journal of Modelling and Simulation*, **42** (4), 653–679.

MALAKOUTI, S. M., GHIASI, A. R., GHAVIFEKR, A. A., AND EMAMI, P. (2022). Predicting wind power generation using machine learning and cnn-lstm approaches. *Wind Engineering*, **46** (6), 1853–1869.

MAPOSA, D., SEIMELA, A. M., SIGAUKE, C., AND COCHRAN, J. J. (2021). Modelling temperature extremes in the limpopo province: Bivariate time-varying threshold excess approach. *Natural Hazards*, **107**, 2227–2246.

MARSAGLIA, G. (2004). Evaluating the normal distribution. *Journal of Statistical Software*, **11**, 1–11.

MASHISHI, D., MAPOSA, D., AND LESAOANA, M. (2020). Comparative analysis of the 100-year return level of the average monthly rainfall for south africa: Parent distribution versus extreme value distributions. *Applied Mathematics and Information Sciences*, **14** (5), 801–807.

MBANDLWA, D. Z. (2023). Cause of the economic decline in south africa. what to expect in the next 10 years from now. *Journal of Survey in Fisheries Sciences*, **10** (2S), 1397–1407.

MORGAN, E., LACKNER, M., VOGEL, R., AND BAISE, L. (2009). Probability distributions for offshore wind speeds. *Elsevier BV*, **52** (1), 15–29.

MUTAVHATSINDI, T., SIGAUKE, C., AND MBUVHA, R. (2020). Forecasting hourly global horizontal solar irradiance in south africa using machine learning models. *IEEE Access*, **8**, 198872–198885.

NASER, M. AND ALAVI, A. (2020). Insights into performance fitness and error metrics for machine learning. *Elsevier BV*.

OKTAVIARINA, A. AND SOFRO, A. (2019). Analysis between temperature and wind speed in east java using bivariate extreme value theory. *Atlantis Press*, **1417** (1), 012019.

PANWAR, N. L., KAUSHIK, S. C., AND KOTHARI, S. (2011). Role of renewable energy sources in environmental protection: A review. *Renewable and Sustainable Energy Reviews*, **15** (3), 1513–1524.

PERERA, F. AND NADEAU, K. (2022). Climate change, fossil-fuel pollution, and children's health. *New England Journal of Medicine*, **386** (24), 2303–2314.

ROLKE, W. AND GONGORA, C. G. (2021). A chi-square goodness-of-fit test for continuous distributions against a known alternative. *Computational Statistics*, **36** (3), 1885–1900.

SAULAT, H., KHAN, M. M., ASLAM, M., CHAWLA, M., RAFIQ, S., ZAFAR, F., KHAN, M. M., BOKHARI, A., JAMIL, F., BHUTTO, A. W., ET AL. (2021). Wind speed pattern data and wind energy potential in pakistan: current status, challenging platforms and innovative prospects. *Environmental Science and Pollution Research*, **28**, 34051–34073.

SEN, P. K. (1968). Estimates of the regression coefficient based on kendall's tau. *Journal of the American Statistical Association*, **63** (324), 1379–1389.

SHAO, B., SONG, D., BIAN, G., AND ZHAO, Y. (2021). Wind speed forecast based on the lstm neural network optimized by the firework algorithm. *Advances in Materials Science and Engineering*, **2021**, 1–13.

SIGAUKE, C. AND BERE, A. (2017). Modelling non-stationary time series using a peaks over threshold distribution with time varying covariates and threshold: An application to peak electricity demand. *Energy*, **119**, 152–166.

SIKHWARI, T., NETHENGWE, N., SIGAUKE, C., AND CHIKOORE, H. (2022). Modelling of extremely high rainfall in limpopo province of south africa. *Climate*, **10** (3), 33.

SINGH, N. P., SINGH, K. P., AND SINGH, U. (1990). Estimation of frechet disribution parameters by joint distribution of'm'extremes. *Statistica*, **50** (1), 59–69.

SIVARAMANAN, S. (2015). Global warming and climate change, causes, impacts and mitigation. *Central Environmental Authority*, **2** (4).

SOMAN, S. S., ZAREIPOUR, H., MALIK, O., AND MANDAL, P. (2010). A review of wind power and wind speed forecasting methods with different time horizons. *North American Power Symposium*, 1–8.

THEIL, H. (1950). A rank invariant method of linear and polynomial regression analysis. i, ii, iii. proceedings of the koninklijke nederlandse akademie wetenschappen. *Indagationes mathematicae*, **53**, 386–392.

WANG, J., SONG, Y., LIU, F., AND HOU, R. (2016). Analysis and application of forecasting models in wind power integration: A review of multi-step-ahead wind speed forecasting models. *Renewable and Sustainable Energy Reviews*, **60**, 960–981.

WANG, Z., HONG, T., AND PIETTE, M. A. (2020). Building thermal load prediction through shallow machine learning and deep learning. *Applied Energy*, **263**, 114683.

WORLD HEALTH ORGANIZATION (2019). Global action plan on physical activity 2018-2030: more active people for a healthier world. *World Health Organization*.

YU, Y., SI, X., HU, C., AND ZHANG, J. (2019). A review of recurrent neural networks: Lstm cells and network architectures. *Neural computation*, **31** (7), 1235–1270.

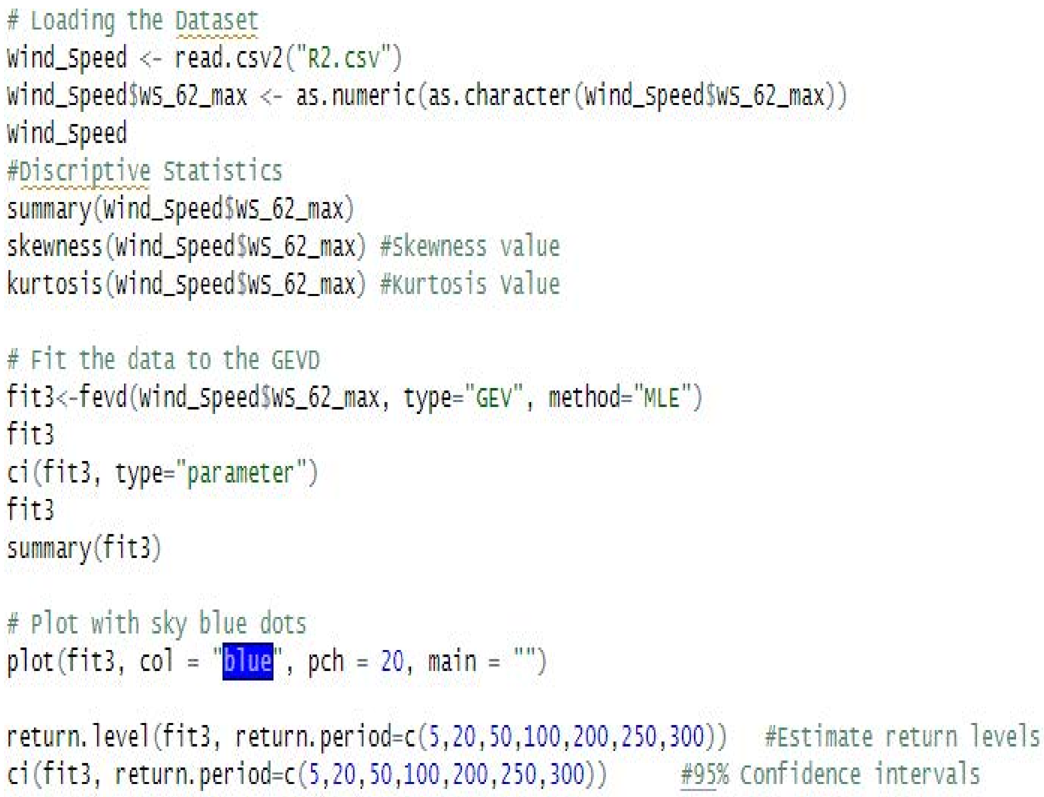
ZAHRAN, B., AYYOUB, B., ABU-AIN, W., HADI, W., AND AL-HAWARY, S. (2023). A fuzzy based model for rainfall prediction. *International Journal of Data and Network Science*, **7** (1), 97–106.

ZENG, X., WANG, D., AND WU, J. (2015). Evaluating the three methods of goodness of fit test for frequency analysis. *Journal of Risk Analysis and Crisis Response*, **5** (3).

ZHANG, J., LU, J., PAN, J., TAN, Y., CHENG, X., AND LI, Y. (2022). Implications of the development and evolution of global wind power industry for china—an empirical analysis is based on public policy. *Energy Reports*, **8**, 205–219.

Appendix

SELECTED R CODES



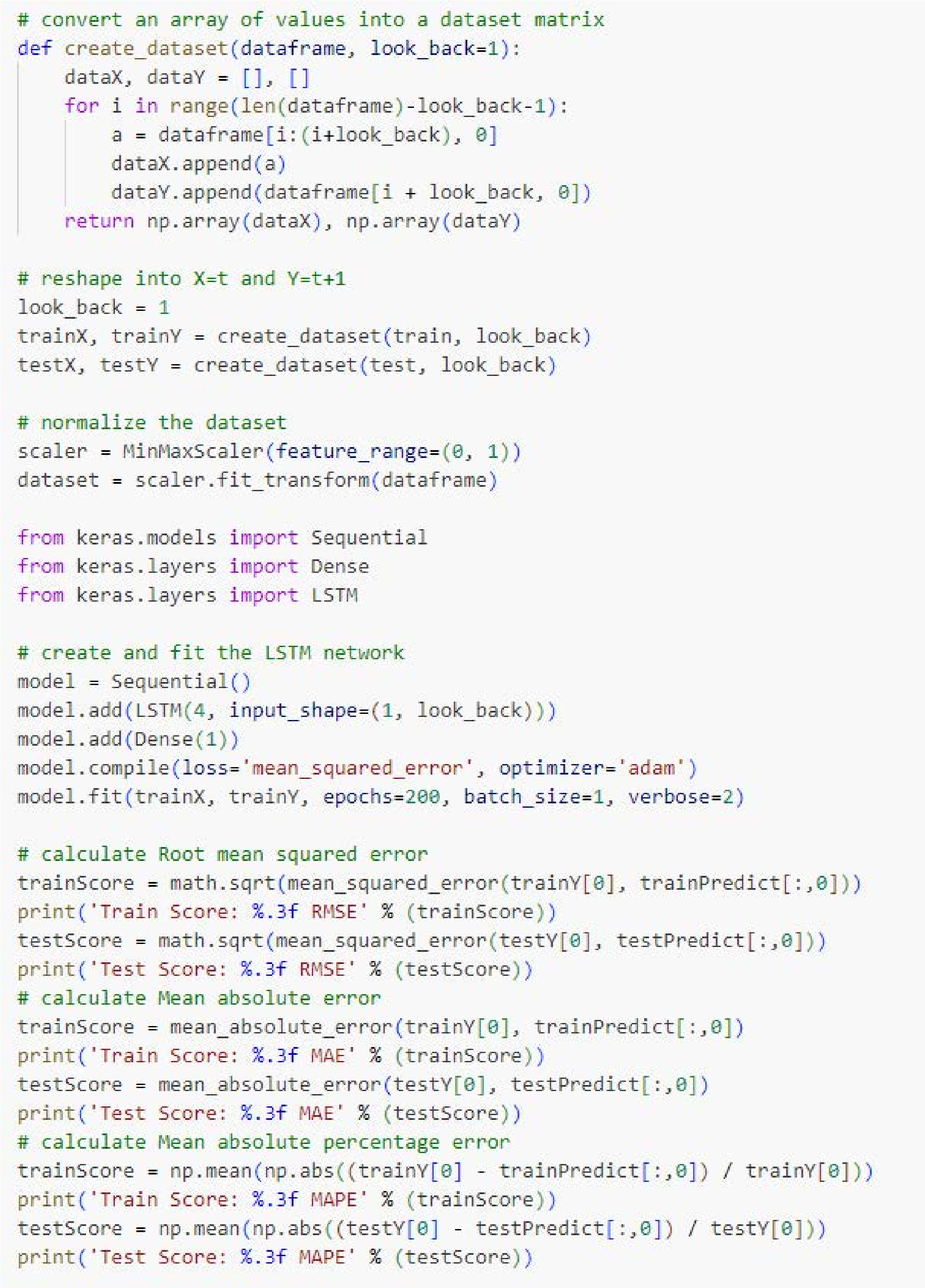
```
# Loading the Dataset
Wind_Speed <- read.csv2("R2.csv")
Wind_Speed$WS_62_max <- as.numeric(as.character(Wind_Speed$WS_62_max))
Wind_Speed
#Discriptive Statistics
summary(Wind_Speed$WS_62_max)
skewness(Wind_Speed$WS_62_max) #Skewness value
kurtosis(Wind_Speed$WS_62_max) #Kurtosis Value

# Fit the data to the GEVD
fit3<-fevd(Wind_Speed$WS_62_max, type="GEV", method="MLE")
fit3
ci(fit3, type="parameter")
fit3
summary(fit3)

# Plot with sky blue dots
plot(fit3, col = "blue", pch = 20, main = "")

return.level(fit3, return.period=c(5,20,50,100,200,250,300))   #Estimate return levels
ci(fit3, return.period=c(5,20,50,100,200,250,300))      #95% Confidence intervals
```

SELECTED PYTHON CODES



```python
# convert an array of values into a dataset matrix
def create_dataset(dataframe, look_back=1):
    dataX, dataY = [], []
    for i in range(len(dataframe)-look_back-1):
        a = dataframe[i:(i+look_back), 0]
        dataX.append(a)
        dataY.append(dataframe[i + look_back, 0])
    return np.array(dataX), np.array(dataY)

# reshape into X=t and Y=t+1
look_back = 1
trainX, trainY = create_dataset(train, look_back)
testX, testY = create_dataset(test, look_back)

# normalize the dataset
scaler = MinMaxScaler(feature_range=(0, 1))
dataset = scaler.fit_transform(dataframe)

from keras.models import Sequential
from keras.layers import Dense
from keras.layers import LSTM

# create and fit the LSTM network
model = Sequential()
model.add(LSTM(4, input_shape=(1, look_back)))
model.add(Dense(1))
model.compile(loss='mean_squared_error', optimizer='adam')
model.fit(trainX, trainY, epochs=200, batch_size=1, verbose=2)

# calculate Root mean squared error
trainScore = math.sqrt(mean_squared_error(trainY[0], trainPredict[:,0]))
print('Train Score: %.3f RMSE' % (trainScore))
testScore = math.sqrt(mean_squared_error(testY[0], testPredict[:,0]))
print('Test Score: %.3f RMSE' % (testScore))
# calculate Mean absolute error
trainScore = mean_absolute_error(trainY[0], trainPredict[:,0])
print('Train Score: %.3f MAE' % (trainScore))
testScore = mean_absolute_error(testY[0], testPredict[:,0])
print('Test Score: %.3f MAE' % (testScore))
# calculate Mean absolute percentage error
trainScore = np.mean(np.abs((trainY[0] - trainPredict[:,0]) / trainY[0]))
print('Train Score: %.3f MAPE' % (trainScore))
testScore = np.mean(np.abs((testY[0] - testPredict[:,0]) / testY[0]))
print('Test Score: %.3f MAPE' % (testScore))
```