PROBABILISTIC RENEWABLE ENERGY MODELLING IN SOUTH AFRICA

By

Thakhani Ravele
11630148

A thesis submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy in STATISTICS

in the

Department of Mathematical and Computational Sciences
Faculty of Science, Engineering and Agriculture
University of Venda, Thohoyandou, Limpopo
South Africa

Supervisor: **Prof Caston Sigauke**

Co-Supervisor: **Dr Lodwell Jhamba**

Abstract

The variability of solar power creates problems in planning and managing power system operations. It is critical to forecast accurately in order to maintain the safety and stability of large-scale integration of solar power into the grid. Accurate forecasting is vital because it prevents transmission obstruction and maintains a power equilibrium. This thesis uses robust models to solve this problem by addressing four main issues. The first issue involves the construction of quantile regression models for forecasting extreme peak electricity demand and determining the optimal number of units to commit at minimal costs for each period using the forecasts obtained from the developed models. The bounded variable mixed-integer linear programming (MILP) model solves the unit commitment (UC) problem. This is based on priority constraints where demand is first met from renewable energy sources followed by energy from fossil fuels. Secondly, the thesis discusses the modelling and prediction of extremely high quantiles of solar power. The methods used are a semi-parametric extremal mixture (SPEM), generalised additive extreme value (GAEV) or quantile regression via asymmetric Laplace distribution (QR-ALD), additive quantile regression with covariate t (AQR-1), additive quantile regression with temperature variable (AQR-2) and penalised cubic regression smoothing spline (benchmark) models. The predictions from this study are valuable to power utility decision-makers and system operators in knowing the maximum possible solar power which can be generated. This helps them make high-risk decisions and regulatory frameworks requiring high-security levels. As far as we know, this is the first application to conduct a comparative analysis of the proposed robust models using South African solar irradiance data. The interaction between global horizontal irradiance (GHI) and temperature helps determine the maximum amount of solar power generated. As temperature increases, GHI increases up to the point that it increases at a decreasing rate and then decreases.

Therefore, system operators need to know the temperature range in which the maximum possible solar power can be generated. The study used the multivariate adaptive regression splines and extreme value theory to determine the maximum temperature to generate the maximum GHI ceteris paribus. Lastly, the study discusses extremal dependence modelling of GHI with temperature and relative humidity (RH) using the conditional multivariate extreme value (CMEV) and copula modes. Due to the nonlinearity and different structure of the dependence on GHI against temperature and RH, unlike previous literature, we use three Archimedean copula functions: Clayton, Frank and Gumbel, to model the dependence structure. This work was then extended by constructing a mixture copula model which combined the Frank and Gumbel models. One of the contributions of this thesis is the construction of additive quantile regression models for forecasting extreme quantiles of electrical load, which are then used in solving the UC problem with bounded MILP with priority constraints. The other contribution is developing a modelling framework that shows that GHI converges to its upper limit if temperature converges to the upper bound. Another contribution is constructing a mixture of some copulas for modelling the extremal dependence of GHI with temperature and RH. This thesis reveals the following key findings:

(*i*) the additive quantile regression model is the best-fitting model for hours 18:00 and 19:00. In contrast, the linear quantile regression model is the best-fitting model for hours 20:00 and 21:00. The UC problem results show that using all the generating units, such as hydroelectric, wind power, concentrated solar power and solar photovoltaic is less costly.

(*ii*) the AQR-2 was the best-fitting model and gave the most accurate prediction of quantiles at $\tau = 0.95, 0.97, 0.99$ and 0.999, while at 0.9999-quantile, the GAEV model had the most accurate predictions.

(*iii*) the marginal increases of GHI converge to 0.12 W/m^2 when temperature converges to 44.26 °C and the marginal increases of GHI converge to -0.1 W/m^2 when RH converges to 103.26%. Conditioning on GHI, the study found that temperature and RH variables have a negative extremal dependence on large values of GHI.

(*iv*) the dependence structure between GHI and variable temperature and RH is asymmetric. Furthermore, the Frank copula is the best-fitting model for variable temperature and RH, implying the presence of extreme co-movements.

The modelling framework discussed in this thesis could be useful to decision-

makers in power utilities, who must optimally integrate highly intermittent renewable energies on the grid. It could be helpful to system operators that face uncertainty in GHI power production due to extreme temperatures and RH, including maintaining the minimum cost by scheduling and dispatching electricity during peak hours when the grid is constrained due to peak load demand.

Keywords: *Bivariate Dependence Modelling, Extremal mixture model, Global Horizontal Irradiance Forecasting, Multivariate Adaptive Regression Spline, Quantile regression, Unit commitment.*

Declaration

I, THAKHANI RAVELE, [11630148], hereby declare that the thesis titled: "Probabilistic renewable energy modelling in South Africa" for the Ph.D. degree in Statistics at the University of Venda, hereby submitted by me, has not been submitted for any degree at this or any other university, that it is my own work in design and in execution, and that all reference material contained therein has been duly acknowledged.

Signed (Student):..... Date: 12 March 2024

Dedication

To

Mashudu Melta Ravele,

Dakalo Ravele

and

Gundolashu Ravele.

Acknowledgment

I express my deepest gratitude to my supervisor Prof Caston Sigauke for his patience, inspiration, guidance and mentoring throughout my studies, for introducing me to the world of statistics of extremes and for co-authoring papers based on this thesis. I thank Dr Lordwell Jhamba, my co-supervisor, for his suggestions, guidance and encouragement to update and improve my writing skills. I want to express my gratitude for the insightful criticism provided by the numerous reviewers who looked through several papers from my thesis destined for publication in peer-reviewed journals. The reviewer's comments widened my perspective on the field of statistics of extremes and significantly raised the quality of this thesis.

I am very grateful to Eskom (South Africa's power utility company), Southern African Universities Radiometric Network and the South African Weather Services for providing the data used in this thesis. I am also greatly indebted to the DSI–CSIR Inter bursary Support (IBS) Programme for financial support. I am grateful to my parents (Mr Fhatuwani and Mrs Mashudu Melta Ravele) and my sister (Ms Marcia Wamashudu Ravele) for their patience, love and financial support. I appreciate them for being the pillars of my strength in encouraging me not to allow people's opinions to become the focus of my meditation. The courageous words of Mrs Dakalo Ravele, Mr Dzulani Mashavhela, Mr Matodzi Mudau, Mrs Fulufhedziso Mudau and my pastor Bishop T.G. Mungoni Muhali, led to the whole idea of furthering my PhD studies. I would like to recognise all staff members and students in the Department of Mathematical and Computational Sciences (DMCS) and, particularly, the Statistics Unit in DMCS at the University of Venda for developing my career. Above all, I want to express my sincere gratitude and thankfulness to the Almighty God for giving me the knowledge, health and strength to start this study project and see it through to the end.

Table of contents

List of Figures

List of Tables

Abbreviations

AQR	Additive quantile regression
AQR-1	Additive quantile regression with covariate t
AQR-2	Additive quantile regression with temperature variable
AIC	Akaike Information criterion
AI	Artificial Intelligence
ALD	Asymmetric Laplace distribution
AVG	Average
AveMaxT	Average of the maximum temperature
AveMinT	Average of the minimum temperature
AveTem	Average daily temperature
BIC	Bayesian Information Criterion
BM	Benchmark model
RC	Cost in Rand
CEM	Conditional extremes model
CSP	Concentrated solar power
CRPS	Continuous rank probability score
CC	Convex combination
CDF	Cumulative distribution function
DSS	Dawid-Sebastiani score
DHI	Diffuse horizontal irradiance
DNI	Direct normal irradiance
E	East
EVT	Extreme value theory
FFNN	Feed-forward neural networks
Fri	Friday
GAEV	Generalised additive extreme value
GAM	Generalised additive model
GEVD	Generalised extreme value distribution

GCV	Generalised cross-validation
GPD	Generalised Pareto distribution
GHI	Global Horizontal Irradiance
LR	Lagrangian relaxation
Lasso	Least absolute shrinkage and selection operator
LQR	Linear quantile regression
LSTM	Long short-term memory networks
Max	Maximum
ML	Maximum likelihood
MED	Median
MW	Megawatt
Min	Minimum
MILP	Mixed-integer linear programming
MARS	Multivariate Adaptive Regression Spline
Mon	Monday
NLQR	Non-linear quantile regression
nonlintrend18	Non-linear trend for hour 18:00
nonlintrend19	Non-linear trend for hour 19:00
nonlintrend20	Non-linear trend for hour 20:00
nonlintrend21	Non-linear trend for hour 21:00
NWP	Numerical Weather Prediction
OLSR	Ordinary least square regression
PSO	Particle swarm optimisation
POT	Peaks over threshold
PV	Photovoltaic
PLF	Pinball loss function
Prod	Production
QR	Quantile regression
QRA	Quantile regression averaging
QR-ALD	Quantile regression via asymmetric Laplace distribution
RH	Relative humidity
RES	Renewable energy sources
SARFIMA	Seasonal autoregressive fractionally integrated moving average
Sat	Saturday
SPEM	Semi-parametric extremal mixture
Soln	Solution
S	South
SAURAN	Southern African Universities Radiometric Network

Std	Standard deviation
Sun	Sunday
SVR	Support vector regression
SDGs	Sustainable Development Goals
Thu	Thursday
TC	Total cost
Tue	Tuesday
UC	Unit commitment
Wed	Wednesday

List of Notation and Special Symbols

F_i	Average production cost of unit i
h	Bandwidth parameter
β_{ij}	Basis coefficients
b_{ij}	Basis functions
$F_t(x_t)$	Bulk model which fit a nonparametric model (kernel density)
$\log f(y;\beta_\theta)$	Censored log-likelihood model
$Q_\tau(Y\|X)$	Conditional quantile of order $\tau \in (0,1)$ of Y given X
$f_{ALD,\tau}$	Density function of asymmetric Laplace distribution
ε_t	t^{th} error term
$F_{t,\mu}(x)$	Excess distribution function over a sufficiently high threshold μ
ξ	Extreme value index (shape parameter)
$q_{Y\|X}(\tau)$	Function for minimising parameter estimates of Additive quantile regression
$\text{GCV}(M)$	Generalised cross-validation function
$G_\xi(x)$	Generalised extreme value distribution function
$H_\xi(x)$	Generalised Pareto distribution function
$y_{t,\mu,\theta}$	Global horizontal irradiance
$\mathbf{I}$	Indicator function
$F^{-1}_{Y\|X}(\tau$	Inverse of conditional distribution function
μ	Location parameter
ℓ	log-likelihood
$\log f(y;\beta)$	Log-likelihood based on fixed τ
L_{AV}	Lower averages model
L_{Md}	Lower median model
$P^{ht}_{Gi}(min)$	Lower limit of the power output
$K(.)$	Kernel function
q_τ	Mapping function

$\hat{\theta}_{t,h,\mu}$	Maximisation of the weighted quasi-log-likelihood estimator
$f(x)$	Multivariate Adaptive Regression Spline function
ln	Natural logarithm
P_{Gi}^{ht}	Northern Cape system load
$G_{\mu,\theta}(x)$	Pareto distribution which approximate the cumulative distribution function $F_{t,\mu}(x)$
$\pi_1(x)$	Penalised cubic regression smoothing spline function
$\ell(\beta_\lambda, \lambda)$	Penalised log-likelihood function
$V(\beta, \gamma, \sigma)$	Penalised pinball loss function
S_λ	Penalty matrix
$\rho_\tau(z)$	Pinball loss function
S_i	Positive semi-definite matrices
P_R^{ht}	Power reserve at hour h on the day t
$\|\mathrm{S}_\lambda\|_+$	Product of positive eigenvalues of matrix S_λ.
QSτ	Quantile score
$\xi_{t,\tau}$	Random error term
$\ell(\beta_\lambda)$	Restricted maximum likelihood function
σ	Scale parameter
$W_{t,h}(t_i)$	Set of weights for $i = 1, ..., n$
s	Smooth function
λ	Smoothing parameters
ρ	Spearman's coefficient of correlation
$b_{ij}(x_{ti})$	Spline basis functions
F_{si}	Start-up cost of unit i at hour h
U_{AV}	Upper averages model
U_{Md}	Upper median model
$P_{Gi}^{ht}(max)$	Upper limit of the unit power output
β_τ	Vector of parameters
γ	Vector of smooth parameters
β_{ij}	i^{th} unknown coefficient (parameter)
$\hat{n}_{t,h,\mu}$	Weighted number of observations above the threshold μ

Research Outputs

A list of research outputs from this thesis is given below.

Peer Reviewed Journal Publications

1. Ravele, T., Sigauke, C. and Jhamba, L. (2022). Economic dispatch of electrical power in South Africa: An application to the Northern Cape province. Statistics, Optimization & Information Computing, Vol. 10 (4), pp. 1235–1249. http://www.iapress.org/index.php/soic/article/view/1057

2. Ravele, T., Sigauke, C. and Jhamba, L. (2022). Estimation of Extreme Quantiles of Global Horizontal Irradiance: A Comparative Analysis Using an Extremal Mixture Model and a Generalised Additive Extreme Value Model. Mathematics and Statistics, Vol. 10 (1), pp. 116–133. https://www.hrpub.org/journals/article_info.php?aid=11656

3. Sigauke, C., Ravele, T. and Jhamba, L. (2022). Extremal Dependence Modelling of Global Horizontal Irradiance with Temperature and Humidity: An Application Using South African Data. Energies, Vol. 15 (16), pp. 5965. https://www.mdpi.com/1996-1073/15/16/5965

Author contributions:

Conceptualization, T.R. and C.S.; methodology, T.R.; software, T.R.; validation, T.R. and C.S.; formal analysis, T.R.; investigation, T.R.; data curation,

T.R. and C.S.; writing—original draft preparation, T.R.; writing—review and editing, T.R., C.S. and L.J.; visualisation, T.R. and C.S.; supervision, C.S. and L.J.; project administration, C.S. and L.J.

International Conferences

1. Ravele, T., Sigauke, C. and Jhamba, L., Extremal Dependence Modelling of Global Horizontal Irradiance with Temperature and Humidity: An Application Using South African Data. 42^{nd} International Symposium on Forecasting, Oxford, England 10^{th} to 13^{th} July 2022, University of Oxford, Exam School.

Local Conferences

1. Ravele, T., Sigauke, C. and Jhamba, L., Estimation of extreme quantiles of GHI: A comparative analysis using an extremal mixture model and a generalised additive extreme value model. 63^{rd} Annual Conference of the South African Statistical Association, 28^{th} November to 3^{th} December 2022, George, Western Cape, South Africa.

2. Ravele, T., Sigauke, C. and Jhamba, L., Economic dispatch of electrical power in South Africa: An application to the Northern Cape province. 64^{th} Annual Conference of the South African Statistical Association, 27^{th} November to 1^{st} December 2023, Durban, Kwazulu-Natal, South Africa.

Data Availability Statement

The data and R-codes used in this thesis can be downloaded from https://github.com/ravelethakhani.

Chapter 1

Introduction

1.1 Introduction

Current considerations from the United Nations focus on transforming the world through implementing Sustainable Development Goals (SDGs) for the 2030 Agenda. Goal 7 of the 17 SDGs aims to ensure access to affordable, reliable, sustainable and modern energy for all. By 2030 goal number 7 envisions achieving five objectives to improve current energy systems, which are to (1) ensure that everyone has access to affordable, dependable and modern energy services, (2) significantly increase the share of renewable energy in the global energy mix, (3) double the rate of improvement in energy efficiency globally and (4) strengthen international cooperation to facilitate access to clean energy research and technology, including renewable energy, energy efficiency, advanced cleaner fossil-fuel technology and promote investment in energy infrastructure (Olaiya, 2016).

Ensuring global access to affordable electricity means considering clean (renewable) energy sources, such as solar, wind and thermal. Decision-makers

in the energy sector require accurate forecasting for uncertain future demand because the energy supply, demand and prices are rising, volatile and unpredictable. Due to running out or lack of fossil fuels and climate change (i.e. global warming, depletion of the ozone layer) with the rising global energy demand, the energy policymakers in developed and developing countries are concerned with "how to meet the 21^{st} century energy demand". Researchers support using and developing renewable energy sources (RES) and solar energy as useful energy sources for producing power for residential, commercial and industrial applications (Hong et al., 2016; Raza et al., 2016; Sobri et al., 2018).

According to Chaturvedi and Isha (2016), a combination of solar energy and accurate forecasting enhances effective planning, energy distribution, and service quality. Solar energy is the reliable RES because the amount of power received by the earth from solar radiation is approximately 8×10^{11} MW which is more than world energy consumption (Shah et al., 2015; Wengenmayr and Bührke, 2011). In the past ten years, solar power generation has risen rapidly in many countries (Alessandrini et al., 2015). South Africa is recently considering more renewable energy technologies into the energy mix. Solar energy is the most viable RES in South Africa because of the high level of solar radiation (Pillay et al., 2014).

This study aims to develop a modelling framework which combines solar power forecasting with optimal grid integration. Understanding peak electricity demand is important for optimal grid integration. Furthermore, this

study considers modelling hourly extreme peak electricity demand at high tail quantiles of the distribution. Accurate forecasts are imperative for solar power to be integrated optimally into the electricity grid. In this thesis, we concentrate on comparative analysis of semi-parametric extremal mixture, generalised additive extreme value or quantile regression via asymmetric Laplace distribution, additive quantile regression, additive quantile regression with temperature variable and penalised cubic regression smoothing spline (benchmark) models for probabilistic forecasting of hourly global horizontal irradiance (GHI) at extremely high quantiles.

Solar power generation can be maximised by knowing the maximum power generated. By utilising the multivariate adaptive regression splines, extreme value theory and copula models, we seek to determine the temperature at which maximum GHI will result. Analysis of GHI at the extreme tails is necessary to understand solar energy's peak power production. Since GHI depends on meteorological variables, we discuss the extremal dependence modelling of GHI with temperature and relative humidity (RH).

1.2 Background

The purpose of electricity load forecasting is to predict hourly, daily, weekly and yearly demand or peak demand (Goude et al., 2013). Peak electricity load forecasting is essential because it assures the availability of enough power supply. Accurate peak electricity load forecasting is important as it provides future forecasts that prevent system failure and power blackouts (Elamin, 2018). The forecasting of electricity loads can be achieved by sta-

tistical and artificial intelligence methods. The extreme peak demand for electricity, however, will be modelled on an hourly basis. In order to achieve a more sustainable energy future, solar energy needs to be explored. For optimal unit commitments (UC), economical operation dispatch and ensuring a national grid's stability, accurate solar power forecasting is useful.

To optimise electricity production and reduce production costs, grid operators require solar irradiance forecasting. Electricity generated by photovoltaic (PV) systems is commensurate to GHI hitting the ground. Solar power generation is primarily affected by meteorological factors such as temperature, humidity, wind pressure, and solar radiation. Figure 1.1 shows a typical daily load profile for South Africa compared to solar power production. The solar power production curve demonstrates an increase in production from hour 08:00 to 14:00 and a drop until hour 20:00. The load demand curve indicates a peak demand at hour 20:00. It is clear that the load demand curve is above the solar power output curve. Other renewable energy sources on the grid will bridge the gap between load demand and solar power output.

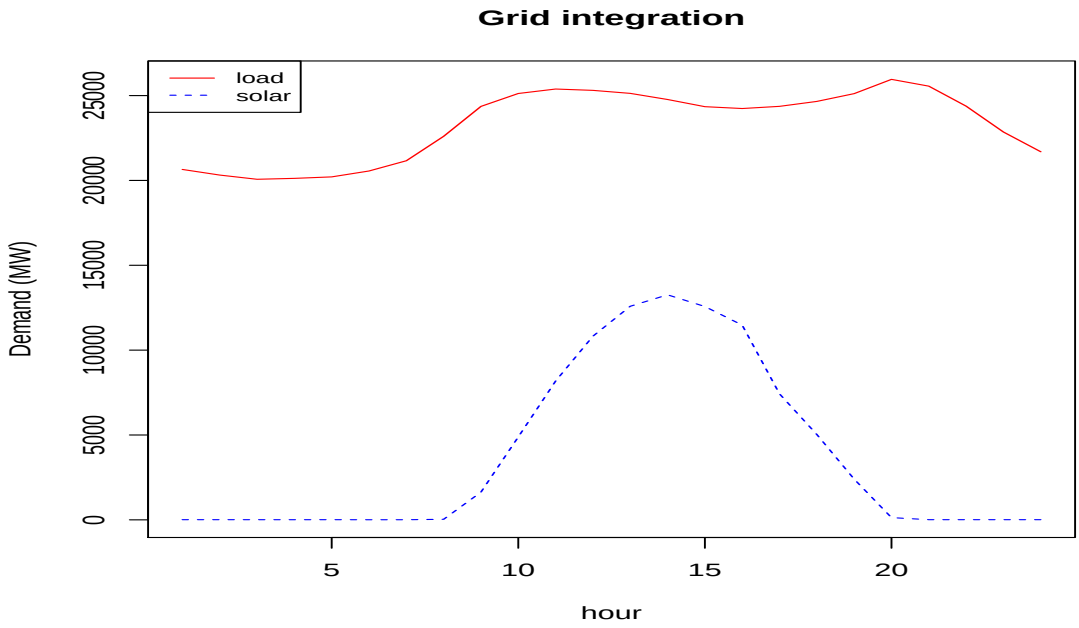

Figure 1.1: Daily load demand profile vs solar power production.

1.3 Statement of the Problem

The variability of solar power creates problems in managing power system planning and operation. Grid integration of solar power requires accurate forecasting to keep the system safe and steady when interconnected. Accurate forecasting is vital because it prevents transmission congestion and maintains a power balance (Nam and Hur, 2018). Forecasting solar power generation (photovoltaic) is advantageous to avoid power generation volume fluctuation, which subsequently impacts power grid planning and operation. Solar power forecasting helps solar power plant designers to design the best solar power plants and is useful in detecting future uncertainty as well as fluctuations of power generation (Sobri et al., 2018; Zhong et al., 2017). Such

solar power fluctuations can damage equipment.

According to Das et al. (2018), there are negative impacts on the electric grid system, such as the creation of instability, unreliability and inconsistencies, as well as complexities of planning the operation, which are caused by the variability of solar power generation. However, to stabilise and secure grid operation with enhanced large-scale solar power integration, accurate forecasting of solar power generation is the key. Significant research has been done on various techniques of forecasting solar power generation (Das et al., 2018), such as clear sky models, regression-based models, Artificial Intelligence (AI) techniques, Numerical Weather Prediction (NWP), remote sensing models, Hybrid systems and Local sensing (Nam and Hur, 2018). Due to the fact that solar power is not planned to be integrated into the grid, it is important to forecast accurately how much solar power can be integrated. Knowing the amount of power needed at a specific time helps the utility, thus maintaining the grid's stability, optimal load scheduling and bidding in electricity markets (Nageem and Jayabarathi, 2017). In this study, we concentrate on an optimal way of integrating the solar power generated into the grid.

1.3.1 Solar power forecasting

Solar power is the most promising energy source for providing the world's energy demands (Fernandez-Jimenez et al., 2012). Solar power forecasting in different time zones is important because it provides the needs of many operations and control activities for the distribution and transmission grids,

like grid regulation, power scheduling, load flow analysis and UC (Inman et al., 2013; Wan et al., 2015). A smart grid requires accurate solar power forecasting to enhance economic operations. Solar power forecasting is done for various time horizons, from short to long-term. Short-term forecasts are done for a few minutes and used for electricity market clearing and solar power storage control. Short-term forecasts are generated for up to 2-3 days ahead, which can be used for various decision-making assignments required in the power system operation and electricity market, including economic dispatch, UC (Inman et al., 2013; Wan et al., 2015), real-time unit dispatching, grid balancing, trading and automatic generation control (Inman et al., 2013).

Medium-term forecasts are produced up to one week ahead, which are important for the scheduling maintenance of solar power plants, transformers, transmission lines and conventional power plants. Long-term forecasts range from months or years and are useful for long-term solar energy assessment, solar power plant planning (Wan et al., 2015), sustainability in utilities and Independent System Operators, scheduling and improving balance area control performance (Inman et al., 2013). Integrating solar power forecasting into the grid increases the standard of the energy supplied to the grid and reduces the weather-dependent ancillary expenses (Inman et al., 2013).

According to Antonanzas et al. (2016), accurate solar power forecasting minimises uncertainties and costs. This prevents possible losses experienced due to deviations between predicted and produced energy for power plant man-

agers. The need for electricity is rising in the world economy nowadays. Worldwide power systems aim to deliver reliable, affordable and sustainable electricity for the future to all. Integrating renewable energy sources (solar power) for electricity generation is important to electrical power systems because it is an effective option for nationwide cheap electricity generation (Zeman, 2014). The modelling framework, which will be discussed, will help guide the planning for the most effective grid integration of solar power.

1.4 Aim and Objectives

1.4.1 Aim

This study aims to develop a modelling framework that combines solar power forecasting with large-scale optimal grid integration.

1.4.2 Objectives

The objectives of the study are to;

(i) Construct quantile regression models for forecasting extreme peak electricity demand and determining the optimal number of units to commit at minimal costs for each period using the forecasts obtained in the developed models.

(ii) Develop and compare models for predicting extremely high quantiles of global horizontal irradiance.

(iii) Determine the maximum temperature that generates the maximum GHI ceteris paribus.

(iv) Estimate the extremal dependence of GHI with temperature and relative humidity (RH) in South Africa.

1.5 Significance of the Study

As a result of this study, a new body of literature on statistical modelling will be added to the existing literature on solar power forecasting and grid integration in South Africa using the latest statistical modelling techniques. The findings of this study will help in the operation of electricity utilities at a minimal cost. Decision-makers from power utilities could benefit from the findings of this study in integrating solar energy efficiently into the electricity grid. Modelling the extremal dependence of GHI against temperature and RH improves the integration of highly intermittent renewable energies into the grid optimally.

1.6 Novelty of the Thesis

The thesis contributes to the existing literature on statistical modelling that combines solar power forecasting with optimal grid integration in South Africa. This thesis contributes the following:

- The study introduces forecasting extremely high quantiles of electricity demand at peak hours. For the hours 18:00 and 19:00, the AQR model appeared to be the most accurate. In contrast, the LQR model appeared to be the most accurate for hours 20:00 and 21:00.The forecasts from the best-proposed models have been used in solving the UC problem by developing a bounded variable mixed integer programming

model in which we minimise the total generation costs with priority constraints. The priority constraints are in such a way that demand is first met from renewable energy sources, followed by energy generated using fossil fuels. It is different from Maswanganyi et al. (2021) study because it uses forecasts from best models to solve UC problem based on extreme high quantiles of electricity demand at peak hours in South Africa.

- Some studies have been done focusing on forecasting GHI using data from South Africa, including others (Chandiwana et al., 2021; Mutavhatsindi et al., 2020). However, to our knowledge, no study has been done on comparative analysis in predicting extremely high hourly GHI data from South Africa. This study closes a gap by comparing the semi-parametric extremal mixture (SPEM), generalised additive extreme value (GAEV) or quantile regression via asymmetric Laplace distribution (QR-ALD), additive quantile regression (AQR-1), additive quantile regression with temperature variable (AQR-2) and penalised cubic regression smoothing spline (benchmark) models in estimating extremely high quantiles of hourly GHI using data from South Africa.

- Accurate predictions of solar power generation are essential for integrating solar power into the grid. The results from this study yield improved forecasts for extremely high quantiles of hourly GHI.

- It is important to know the maximum temperature at which GHI will be generated, at the maximum. This study presents MARS, extreme value theory, and copula models for addressing such issues.

- The study also discusses extremal dependence modelling of GHI with temperature and RH. Empirical results show that the marginal increases of GHI converge to 0.12 W/m^2 when temperature converges to 44.26 °C and the marginal increases of GHI converge to -0.1 W/m^2 when RH converges to 103.26%. Conditioning on GHI, the study found that temperature and RH variables have a negative extremal dependence on large values of GHI.

- The dependence on GHI against temperature and RH is nonlinear and complex. This study proposes three Archimedean copula methods: Clayton, Frank, and Gumbel, which are useful for modelling dependence structures.

1.7 Thesis Structure

The rest of the study is arranged as follows: In Chapter 2, we review existing literature on load forecasting, solar irradiance forecasting, and the summary of studies on extremal dependence modelling of GHI. Chapter 3 provides a general theory of the statistical methods used in this study. Chapters 4, 5 and 6 focus on data modelling, estimations and analytical findings, where Chapter 4 presents the economic dispatch of electrical power in South Africa, Chapter 5 provides the estimation of extreme quantiles of GHI, and Chapter 6 reports details of extremal dependence modelling of GHI with temperature and RH. The study's conclusion, concluding remarks, and directions for future research are described in Chapter 7.

Chapter 2

Literature Review

2.1 Introduction

Solar power forecasting is very crucial for the optimal operation of the solar power uncertainties occurring in the electricity sector. Integrating solar power into an electrical network helps the decision makers plan the production and unpredictable electricity demand (Voyant et al., 2017). Accurate solar power forecasting also involves optimal planning, which minimises the effects of solar power uncertainty, increases the power system's penetration level, improves the system's stability and ensures that additional devices require less maintenance. Therefore, it is very important to produce accurate forecasts that help schedule solar power plants where the power generation will closely match the predicted load profile (Sobri et al., 2018). Factors such as solar irradiance, wind speed, cloud cover, temperature and seasonal variations have significant effects on solar power, which results in uncertain and unpredictable solar power patterns. However, it is important to create a robust, adaptive and intelligent forecast model that accommodates the factors influencing solar power to improve forecast accuracy (Raza et al., 2016).

Sobri et al. (2018) presents some of the methods applied in forecasting solar power production. The methods are categorised into statistical, machine learning, physical and ensemble. The statistical methods consist of autoregressive and regression-based models, while machine learning consists of support vector regression, artificial neural networks and stochastic gradient boosting, among others. Physical methods comprise Sky Imagery, Satellite-Imaging and Numerical Weather Prediction models. The physical models influence the interplay between the dynamic motion and the physical state of solar radiation in the atmosphere, whereas ensemble methods consist of a combination of physical and statistical methods (Sobri et al., 2018). In the past ten years Raza et al. (2016), indicated that world solar power capacity increased exponentially. Due to such growth, solar power plants are increasingly being integrated into the power system.

This chapter is presented as follows. Section 2.2 presents an overview of the literature on load forecasting. A review of literature on solar irradiance forecasting is discussed in Section 2.3, and a review of articles on extremal dependence modelling of GHI is reviewed in Section 2.4. Contributions to the literature and concluding remarks are given in Sections 2.5 and 2.6, respectively.

2.2 An Overview of the Literature on Load Forecasting

2.2.1 Electricity demand forecasting

In recent years, academics and industrialists have paid considerable attention to electricity load forecasting (Goude et al., 2013). Various forecasting techniques have been developed for forecasting electricity loads, including statistical and artificial intelligence methods (Hong and Fan, 2016). Forecasting electricity loads can take various forms depending on what results are desired. The study may be spatial if it focuses primarily on future patterns in a particular region or country. It is considered temporal if the forecast is hourly, daily, monthly or annual. Electricity load forecasting is carried out according to their time horizons, such as very short-term load forecasts, short-term load forecasts, which range between an hour and a week, medium-term load forecasts are generated from a week to a year, and long-term load forecasts range up for more than a year (Hong and Fan, 2016; Feinberg and Genethliou, 2005; Hyndman and Fan, 2009).

In order to prevent system failure and power outages, Elamin (2018) developed a quantile regression model. The goal was to avoid underestimation. The actual demand was compared with estimates from the quantile regression model to investigate the ability of the model to avoid power blackouts (avoid underprediction). The results showed that the daily peak demand upper limit prediction is accurate when using the 1.00 quantile from the 0.97 to 0.99 quantiles of the distribution (Elamin, 2018).

The rising peak electricity load is a consequence of climate change, technological development, and energy policies. Therefore, unit sectors require accurate peak electricity load forecasting. When creating policies and planning power generation, companies and decision-makers rely on accurate energy demand forecasts (Kim et al., 2019). Accurate forecasting of peak electricity demand is essential for planning capacity enlargement and medium-term risk assessment in the electricity sector (Maswanganyi et al., 2017). Peak electricity load forecasting and modelling using South African data are discussed in the literature; see (Maswanganyi et al., 2017; Sigauke and Chikobvu, 2012; Sigauke et al., 2013).

Sigauke and Chikobvu (2012) used the additive model that allows non-linear and nonparametric terms to forecast daily winter peak electricity demand in South Africa. The study revealed that South Africa's peak electricity demand is far more sensitive to cold temperatures than heat. In order to quantify the electricity amount that can be switched from the grid to off-peak times, it is useful to model extreme peak electricity load (Sigauke et al., 2013). Sigauke et al. (2013) modelled increases in peak extreme daily electricity in South Africa, focusing on the distribution tail quantiles of daily peak electricity demand. The forecasting of peak electricity demand in South Africa is studied by Maswanganyi et al. (2017) in which the authors focused on an application of partially linear additive quantile regression models for modelling and forecasting peak electricity demand. Short-term hourly load forecasts were made using additive quantile regression (AQR) models in South Africa. Based on

pinball loss and quantile regression averaging (QRA), forecasts were combined from four developed models. The study found that the AQR model with interactions produced accurate forecasts compared to the QRA model (Sigauke et al., 2018).

2.2.2 Unit commitment

The unit commitment (UC), also known as an optimisation problem, allows the management of generating units (whether they should be on or off) under various restrictions and environments (Saravanan et al., 2013). UC maximises power systems operations with the minimal costs of production and reserve requirements (Dai et al., 2015; Wright, 2013). A reasonable solution to the optimisation problem is very important because it provides operational planners with the optimal number of required generators. According to Hong et al. (2016), there are different methods for determining a reasonable solution to the optimisation problem. Some optimisation techniques used in solving the unit commitment are, Lagrangian relaxation (LR), mixed-integer linear programming (MILP), tabu search, dynamic programming and stochastic programming techniques, see (Saravanan et al., 2013).

However, Kurban and Filik (2008) revealed that many authors used mixed integer programming and LR methods. They developed an approach that integrates short-term load forecasting with UC, which minimises production costs from one thermal plant in the Kutahya region of Turkey. Applying the LR method Kurban and Filik (2008) provided solutions to a UC problem using forecasts from two developed models for electricity demand. Accurate

load forecasts are essential as they are some of the inputs in solving the UC problem.

2.3 A Review of the Solar Irradiance Forecasting

Operators rely on accurate forecasts to maintain uncertainty created by increasing solar penetration (Antonanzas et al., 2016). According to Xie et al. (2018), accurate forecasting of PV power is required for electric power system operation while installing grid-connected PV generation on a large scale. The application of PV power forecasting is used in decision-making for grid dispatch and to complement the control of other multiple power sources. Potential statistical modelling errors, including changing weather patterns, pose challenges in forecasting solar power accurately (Azimi et al., 2016). To reduce the intermittent solar power and operational expenses when integrating it into the electricity grid, accurate forecasts are important and capture the uncertainty surrounding the forecasts (Ghayekhloo et al., 2015).

Reliable, accurate forecasts are required to maintain the balance of fluctuating PV energy generation and make the integration to render better quality service (Diagne et al., 2013). The forecasts for solar power are mainly classified into very short, short, medium and long-term categories (Kostylev et al., 2011). Kostylev et al. (2011) defined very short-term forecasts as 15 minutes to a few hours ahead with a granularity of 30 seconds to 5 minutes for events ramp up and variability. For short-term forecasts, 24 to 168 hours ahead with a granularity of hourly. These forecasts are useful for transmission schedul-

ing, including unit commitment. Medium-term forecasts are important for planning and asset optimisation, ranging from a week to a year ahead. Long-term forecasts are generated from a year up to several years ahead and are used for planning the extension of the electricity grid, including the building of new power plants (Kostylev et al., 2011).

According to Lago et al. (2018) and Voyant et al. (2017), solar irradiance short-term forecasts are essential for operational planning, including scheduling of power systems. The forecasting techniques of solar irradiance are typically separated into physical and statistical approaches, including machine learning algorithms (Diagne et al., 2013; Mutavhatsindi et al., 2020). Physical approaches use numerical weather predictions, and statistical approaches use historical time solar time series data (Mpfumali et al., 2019; Sun et al., 2018; Diagne et al., 2013; Mutavhatsindi et al., 2020). There are three types of solar irradiance, which are direct normal irradiance (DNI), diffuse horizontal irradiance (DHI) and global horizontal irradiance (GHI), respectively (Lago et al., 2018). GHI is the sum of DNI and DHI.

Solar energy plays an important role in electricity markets (Antonanzas et al., 2016). Antonanzas et al. (2016) compiled a review of solar power forecasting focusing on the current improvements and future projections of solar energy. The idea of the study was to present a precise forecast with its analysis of the economic implications. The study revealed that most research focused on the day-ahead forecast horizon. Utilities, independent system operators, solar power producers and energy traders require accurate solar power forecasts

for core operations (Kostylev et al., 2011). Kostylev et al. (2011) proposed a set of standards for performance evaluation of solar power forecasts. The standards consist of sound methodologies and a lot of field experience, which provide accurate inter-agency comparisons of forecast performance. A hybrid short-term forecasting model was proposed by Xie et al. (2018) to solve the problem of forecasting accuracy using solar data from the Yunnan PV power generating plant. Several forecast evaluation metrics were used, which include normalised absolute average error and normalised root mean square error, among others. Results from the study revealed that the proposed model was the best-fitting model.

Solar irradiance forecasting is discussed in the literature using South African data. In a recent study, Mutavhatsindi et al. (2020) predicted hourly GHI using the University of Pretoria radiometric station data. A comparative analysis of long-short term memory (LSTM), feed-forward neural networks (FFNN) and support vector regression (SVR) models were done, with results showing that the FFNN model gives the most accurate forecasts. The individual models' forecasts were combined to get an ensemble LSTM-SVR-FFNN model via quantile regression averaging and convex combination (CC). The ensemble LSTM-SVR-FFNN model via QRA was the best-fitting model with the most accurate forecasts. The study suggested that good and thorough metrics for evaluation and statistical tests render more perspective into the proposed forecasting methods (Mutavhatsindi et al., 2020). Using data from the Tellerie radiometric station in the Northern Cape province of South Africa, Mpfumali et al. (2019) developed partially linear AQR models, which

were then used in forecasting day-ahead GHI. The variables were selected via hierarchical interactions using the least absolute shrinkage and selection operator (Lasso). The individual model forecasts were then combined using QRA and CC. The study revealed that QRA yields the most accurate forecasts (Mpfumali et al., 2019).

Adeala et al. (2015) applied a multiple linear regression model to forecast global solar irradiance in South Africa. The covariates used in the study were weather variables, traditional extraterrestrial irradiance and sunshine hours. The study showed that the solar irradiance models' performance and accuracy in other areas improved by including weather parameters (Adeala et al., 2015). Ranganai and Sigauke (2020) solved the issue of long-range dependence ingrained in the solar irradiance data in South Africa using three models, which are seasonal autoregressive fractionally integrated moving average (SARFIMA), harmonically Coupled SARIMA and regression model with SARFIMA error terms. In the study, AQR was used as a benchmark model. Empirical results from the study suggest that long memory is anti-persistent in all the developed models (Ranganai and Sigauke, 2020).

2.4 Review of Articles on Extremal Dependence Modelling of GHI

Although several studies have used South African data in modelling and forecasting GHI, none, to the best of our knowledge, has used conditional multivariate extreme value theory together with copula functions in modelling the dependence and correlation of GHI with temperature, including

GHI with relative humidity.

2.4.1 Multivariate Adaptive Regression Spline models

The Multivariate Adaptive Regression Spline model (MARS) is a nonparametric and improved version of the traditional linear regression model that maintains the clarity of the conventional multiple linear regression. Due to its nonparametric property, the MARS model provides reliable results in training and forecasting. Due to its simplicity and reasonable interpretation, the MARS model was also suggested as a feasible alternative by Li et al. (2016). A related study by Li et al. (2019) suggested a model for predicting hourly global solar radiation on a horizontal surface, especially for areas with restricted measured meteorological parameters using the MARS model.

The MARS models have not been used to analyse and forecast GHI data to our knowledge. However, they have been used for modelling electricity demand; see the study by Sigauke and Chikobvu (2010). They developed a MARS model for predicting daily electricity peak demand. The study concluded that the MARS model is useful to management and performed well compared to the piecewise linear regression model. The MARS method produces easily understood results by management, making it an excellent tool for handling high-dimensional situations with the structure of the complex model, such as interactions, nonlinearity and missing data (Sigauke and Chikobvu, 2010).

2.4.2 Conditional Multivariate Extreme Value models

Copulas, joint tail models and other multivariate extreme approaches assume that all the variables will be large simultaneously. The conditional extremes model (CEM) overcomes this limitation (Heffernan and Tawn, 2004; Keef et al., 2013; Tilloy et al., 2020). When studying the relationship between several variables, conditional extremes modelling is crucial (Maposa et al., 2021). Maposa et al. (2021) used a time-varying threshold and bivariate conditional extremes model to forecast the extremal dependence of monthly maximum temperature from four meteorological stations of Limpopo province, South Africa. The results showed significant negative and positive extremal dependence in some meteorological stations.

Gonçalves et al. (2021) estimated extremely low (below 0.05-quantiles) and extremely high (above 0.95-quantiles) quantiles of wind power using wind power data from a plant situated in Galicia, Spain. The study applied conditional extreme value theory estimators by integrating gradient boosting trees and truncated generalised Pareto distribution. The proposed method outperformed benchmark models based on pinball losses. Another study by Nemukula et al. (2018) used the conditional extreme value modelling approach in modelling the dependence of extreme temperatures from three meteorological stations in the province of Limpopo, South Africa. The authors argued that since the main driver of electricity demand is temperature, extremely high temperatures impact electricity generation to meet demand.

2.4.3 Bivariate Dependence Modelling Using Copulas

Copulas are used in modelling the correlation and dependence between several variables based on their marginal distributions (Nelson, 2006). There are several copula models, and among them are the Archimedean, extreme value, Gaussian and Student's-t copulas. In this thesis, we focus on the Archimedean copulas. Archimedean copulas are a special type of copulas. Their properties have enabled them to be used in various areas because they are easy to construct (Nelson, 2006). Yet and Masseran (2021) fitted the Gumbel, Joe, Frank, Clayton, Gaussian and Student's-t copula functions to evaluate the dependence between the sky clearness index and solar radiation. Based on the Akaike Information Criterion (AIC), the Gaussian copula was best used to model the extremal dependence between solar radiation and the sky clearness index. In Munkhammar and Widén (2016), a copula-based model was used to examine the relationship between the beam radiation and diffuse radiation at different locations using data from Sweden. Another study by Bazrafshan et al. (2015) used five copula functions to model the dependence structure of sunshine duration hours with solar radiation at nine radiometric stations in Iran. Based on the two information criteria, AIC and Schwartz information criterion, the Clayton copula yielded the best performance.

Some studies have been carried out to correlate temperature with global solar radiation. These include a study by Prieto et al. (2009), which estimated the correlation between global solar irradiation and temperature. The proposed method was applied to eight radiometric stations in Asturias, a re-

gion northwest of Spain. It was noted from the study that the proposed method was able to explicitly show the influence of the distance to the sea, altitude and reference temperature. In a related study, Munkhammar and Widén (2016) used a two-variable bivariate copula-based method in describing the dependency between the beam and diffuse irradiance data. It was noted from the study that the developed modelling framework significantly simplifies calculations for solar engineering applications.

2.5 Contributions to the Literature

Forecasting the peak electricity load demand within extreme conditions is the solution to the uncertainty of the demand. Traditionally, electricity demand forecasting models have considered mainly average changes. Extending our thinking from the average to the extreme can provide valuable information. This study model hourly peak electricity demand at high tail quantiles of the distribution to close the existing modelling gaps.

In order to maintain the balance of solar energy generation fluctuations, accurate forecasts are required, and integration is necessary to deliver higher-quality services. The role of solar energy in electric markets has transformed forecasting into a crucial skill. Many researchers have been interested in short-term solar power forecasting with hourly cycles (Xie et al., 2018). Based on the literature, no studies have focused on forecasting extremely high quantiles of GHI using South African data. Despite this, this study includes temperature variables to improve the precision of models and use different models for hourly GHI forecasting at extremely high quantiles of

distribution tails.

There have been some earlier studies on forecasting GHI, although most concentrate on forecasting the means or averages of the distribution. While several studies have used South African data in modelling and forecasting GHI, none, to the best of our knowledge, has used conditional multivariate extreme value theory together with copula functions in modelling the dependence and correlation of GHI with temperature, including GHI with relative humidity. Understanding solar energy's peak power production requires analysis of GHI at the extreme tails. The weather variables, temperature and RH, which are the main drivers of GHI, are used to improve accuracy. To the best of the authors' knowledge, none have used MARS models, conditional multivariate extreme value models and Archimedean copulas such as Clayton, Frank and Gumbel to model the extremal dependency of GHI with temperature and RH using South African data before this study. This is the gap that this study seeks to cover.

2.6 Concluding Remarks

Modelling the extreme quantiles of electricity load and solar power generation has been reviewed in this Chapter. While the literature contains few studies on modelling extreme quantiles, no studies have been done on applications of extreme quantiles modelling on solar power generation. On the other hand, this study uses multivariate regression splines, extreme value theory and copula model to solve the extremal dependence of GHI with temperature and humidity. Forecasting solar power generation can help the decision-makers

in the power utility to plan the integration of solar power into the grid while solving the UC problem.

Chapter 3

Methodology

3.1 Introduction

Accurate forecasting methods of electricity demand, including solar power, are important to systems operators, including decision-makers in power utilities, for dispatching electrical energy, including integrating renewable energies onto the grid on a large scale. This chapter discusses the methods used in this thesis: quantile regression (QR)-based methods, extreme value theory (EVT) and copula models. We also discuss the bounded variable mixed integer linear programming (MILP) method and particle swarm optimisation, which are applied to the unit commitment (UC) problem.

The modelling presented in this chapter uses various R statistical packages, including "qgam" (Fasiolo et al., 2020b) and "quantreg" (Koenker et al., 2018) for QR-based models. The EVT aspect will be modelled using R-packages such as "extremefit" developed by Durrieu et al. (2016), "evgam" developed by Youngman (2020),"earth" developed by Milborrow (2018) and "evmix" developed by Hu and Scarrott (2018). UC problem programming

formulation will be solved using Lingo version 18 (Gupta and Ali, 2021). Parameter estimation is done using maximum likelihood (ML) and some of its variants, such as restricted ML.

3.2 Quantile Regression

Quantile regression (QR) was initially introduced by Koenker and Bassett (1978). It provides a systematic approach to studying how covariates affect the response variable of the distribution. QR is a statistical approach which provides prediction models of conditional quantile distribution (Koenker, 2005). It is an extension of ordinary least square regression (OLSR). OLSR is a statistical method for examining the connection between a response variable and one or more independent variables. Instead of the conditional mean established by OLSR, it establishes the conditional function of a dependent variable at different quantiles. QR is a most viable statistical technique because it does not have limitations on conditional mean as it offers location, shape and scale of the whole conditional distribution of the response variable given a set of explanatory variables (Davino et al., 2013). In addition, QR is powerful in predicting outliers (Koenker, 2005).

Let Y denote random variables (i.e. solar power generation) and $\mathbf{X}$ a set of independent variables (i.e. temperature, time, among others). Then the conditional quantile Q_τ of order $\tau \in (0,1)$ of Y given $\mathbf{X}$ is defined as the inverse of conditional distribution function $F_{Y|X}$ of Y knowing $\mathbf{X}$ (Davino et al., 2013):

$$Q_\tau(Y|\mathbf{X}) = F^{-1}_{Y|\mathbf{X}}(\tau) = \inf\{y \in \mathbb{R} : F_{Y|\mathbf{X}}(y) \geq \tau\}. \tag{3.2.1}$$

3.2.1 Linear quantile regression model

Linear quantile regression assumes a linear relationship between the response variable and a vector of explanatory variables when estimating the quantiles of the cumulative distribution function of the response variable (Koenker and Bassett, 1978). Suppose y_t denotes the response variable on day $t = 1, \cdots, n$ at quantile τ^{th} of the regression given in Equation (3.2.2):

$$y_t = \mathbf{X}\beta + \varepsilon_t \tag{3.2.2}$$

and the design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ of the τ^{th} row denoted by $\mathbf{X}$, with β representing a vector of parameters and ε_t denotes the error term. The τ^{th} regression quantile is the solution to the minimisation problem given in Equation (3.2.3).

$$\hat{\beta}(\tau) = \min_{\beta \in \mathbb{R}^p} \left[\sum_{\varepsilon_t \geq 0} \tau |\varepsilon_t| + \sum_{\varepsilon_t < 0} (1 - \tau) |\varepsilon_t| \right], \tag{3.2.3}$$

where $\varepsilon_t = y_t - \mathbf{X}\beta$ denotes the residuals and the τ^{th} regression quantile is denoted by $\hat{\beta}(\tau)$ with $\tau \in (0, 1)$.

Strengths

When dealing with data that can be separated linearly, linear quantile regression works remarkably well. It is simpler to implement, understand and train on.

Limitations

The dependent and independent variables are assumed to be linearly related. Overfitting is a common problem with it.

3.2.2 Non-linear quantile regression model

Non-linear quantile regression assumes a non-linear relationship between the dependent variable and a vector of independent variables when predicting the quantiles of the distribution. In non-linear quantile regression modelling, a non-linear mapping function q_τ converts the vectors of explanatory variables X_t into a viable higher dimensional feature space (Hwang and Shim, 2005). A non-linear quantile regression model is defined in (Hwang and Shim, 2005; Conde Amboage et al., 2018; Koenker, 2017):

$$y_{t,h,\tau} = q_\tau\big(X_t, \beta_\tau\big) + \xi_{t,\tau}; \quad \tau \,\epsilon\, (0,1), \tag{3.2.4}$$

where $y_{t,h,\tau}$ is the response variable on day $t = 1,..,n$ at hour h at quantile τ, β_τ denotes a vector of parameters at quantile τ, q_τ is the mapping function and $\xi_{t,\tau}$ denotes random error term. Equation (3.2.4) can be estimated by:

$$\hat{\beta}(\tau) = \arg\min_{\beta\epsilon\mathbb{R}} \sum_{t=1}^{n} \rho_\tau\big(y_{t,h,\tau} - q_\tau(X_t, \beta)\big), \tag{3.2.5}$$

Advantages

The models' advantage is their flexibility in estimating the conditional distribution of the relevant variable. Based on a big sample approximation, the inference is carried out. The response variable is resilient to outliers.

Limitations

Compared to generalised regression, parameters are more difficult to estimate. The smoothing parameter must be specified. Quantile curves close to outliers tend to be affected by them, i.e. extremal quantiles.

3.2.3 Additive quantile regression model

Brief history of generalised additive model

The generalised additive model (GAM) assumes that independent and dependent variables have additive relationships. GAMs were introduced by Hastie and Tibshirani (1990). GAMs do not assume a functional relationship apriori and, as such, allow for more flexibility in the modelling. A GAM is given as (Sigauke, 2017):

$$y_{t,h,\tau} = \beta_{0h} + \sum_{j=1}^{p} s_{hj}(x_{thj}) + \varepsilon_{th} \tag{3.2.6}$$

where β_{0h} denotes a constant parameter, s_{hj} is a smooth functions and ε_{th} represent error terms that are assumed to be autocorrelated, y_{th} represents response variable, where p denotes the number of variables. The smooth function, s_j can be described as the sum of regression coefficients β_i and basis functions $b_i(x)$, and is given by:

$$s_j(x) = \sum_{i=1}^{q} \beta_i b_i(x) \tag{3.2.7}$$

where q denotes the basis dimension. The smooth effects are estimated using penalised cubic splines, and the function is given by:

$$min_{s_j}[(y_t - \beta_0 - \sum_{j=1}^{p} s_j(x_{tj}))^2 + \sum_{j=1}^{p} \lambda_j(\int (f''(x))^2 dx)], \tag{3.2.8}$$

where $\bigwedge = (\lambda_1, ..., \lambda_p)$ are the penalty parameters which determines the degree of smoothness and control how closely the function estimate resembles the data.

Additive quantile regression model

An additive quantile regression (AQR) model is a hybrid model that integrates GAM and QR models. AQR models were initially used by Gaillard et al. (2016) to estimate short-term load demand and extended by (Fasiolo et al., 2017, 2020a). The AQR model is written as (Gaillard et al., 2016; Fasiolo et al., 2017, 2020a,b):

$$y_{t,h,\tau} = \sum_{i=1}^{p} s_{i,h,\tau}(x_{ti}) \; + \; \varepsilon_{t,h,\tau}; \quad \tau \; \in \; (0,1), \tag{3.2.9}$$

The smooth function, s, is given by:

$$s_i(x) = \sum_{j=1}^{q} \beta_{ij} b_{ij}(x_{ti}), \tag{3.2.10}$$

where β_{ij} represents the i^{th} unknown coefficient (parameter) and $b_{ij}(x_{ti})$ are known as spline basis functions. The parameter estimates of Equation (3.2.9) are obtained by minimising the function given by:

$$q_{Y|X}(\tau) = \sum_{t=1}^{n} \rho_\tau \left(y_{t,\tau} \; - \; \sum_{i=1}^{p} s_{i,\tau}(x_{ti}) \right), \tag{3.2.11}$$

where

$$\rho_\tau(z) = (\tau - 1) z \mathbf{I}(z < 0) + \tau z \mathbf{I}(z \geq 0)$$

represents the pinball loss function. This study uses the penalised pinball loss function discussed in (Fasiolo et al., 2020a). Let $\mu(x_t) = X_t^T \beta$, with x_t denoting the t^{th} row of the $n \times d$ design matrix X. The penalised pinball loss function is then defined as (Fasiolo et al., 2020a):

$$V(\beta, \gamma, \sigma) = \sum_{t=1}^{n} \frac{1}{\sigma} \rho_\tau \left[y_t, \tau - \mu(x_t) \right] + \frac{1}{2} \sum_{i=1}^{p} \gamma_i \beta^T S_i \beta, \tag{3.2.12}$$

where γ denotes a vector of smooth parameters, i.e. $\gamma = \{\gamma_1, ..., \gamma_p\}$, S_i represents the positive semi-definite matrices which are meant to penalise the wiggliness of $\mu(x)$ (Fasiolo et al., 2020a). The term $\frac{1}{\sigma}$ represents the learning rate used to control the weight of the loss and the penalty. For more details, see (Fasiolo et al., 2020a).

Strengths

AQR is a hybrid model which combines GAMS and QR. Distribution estimation is free. The response variable is robust to outliers. Utilise crossover and mutation operators to conduct a global search and explore the search space. It is useful for noisy environments. It is useful in mixed discrete/continuous problems.

Limitations

A covariate smoothing function is necessary for the AQR model. It is challenging to estimate parameters. The size of a high degree of potential exceedances is not specified. Computationally expensive. Implementation is still an art.

3.3 Extreme Value Theory

3.3.1 Brief history of Extreme value theory

EVT is a useful and sufficiently robust modelling framework for the extreme tails behaviour of the heavy-tailed distributions (Gençay and Selçuk, 2004). EVT renders simple parametric models to capture the extreme tails and estimate the risk behaviour of a distribution (Singh et al., 2013). The greatest

concern of EVT is to produce an accurate prediction model of the tails for the distributions, which can be used to forecast the probabilistic behaviour of rare events. EVT provides an asymptotic extrapolation approach to predict data beyond the range (Winter, 2016). The application of EVT consists of two approaches: block maxima and peaks over threshold (POT). Coles (2001); Beirlant et al. (2004, 2006); Dey and Yan (2016); among others provide an overview of EVT. The following are the EVT distributions.

Generalised extreme value distribution

Generalised extreme value distribution (GEVD) is the foundation of EVT, which starts as the asymptotic distribution of sample maximum of independent and identically distributed random variables when the sample size is large (Dey and Yan, 2016). GEVD is known as block maxima, or minima approach, a family of parametric distributions which are the Fréchet, Gumbel and Weibull distributions. Distributions of extreme values are also known as type I, II, and III distributions, respectively. Each type of extreme value distribution corresponds to a limiting distribution of block maxima from various classes of underlying distributions. For instance, distributions with exponentially decreasing tails, such as the normal distribution, fall under type I. However, polynomially decreasing tails, such as Student t distribution, fall under type II. Distributions with finite tails, such as the beta distribution, fall under type III (Hor et al., 2008). The Fisher and Tippett (1928) theorem indicates that the asymptotic distribution of maxima belongs to one of the extreme value distributions without considering the distribution of the observed data. The GEVD model, which combines extreme value distributions

with one parameter presentation (Coles, 2001; Hor et al., 2008) is given by:

$$G_\xi(x) = \exp\left\{ -\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}, \text{if } 1+\xi\left(\frac{x-\mu}{\sigma}\right) > 0,\ \xi \neq 0, \quad (3.3.1)$$

where ξ is the extreme value index (shape parameter), μ is a location parameter and σ denotes a scale parameter. For ξ=0 (i.e. as ξ approaches zero), the distribution is the Gumbel class which is defined by:

$$G_\xi(x) = \exp\left\{ -\exp\left[-\frac{x-\mu}{\sigma}\right]\right\},\ \xi = 0. \quad (3.3.2)$$

When ξ >0, we have the Fréchet class of distribution, and when ξ <0, GEVD corresponds to the Weibull class of distributions. Quantiles are commonly used for applications involving extreme values. We estimate the GEVD's extreme quantiles as follows:

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}\{-\log(1-p)\}^{-\xi}, & \text{for } \xi \neq 0, \\ \mu - \sigma\log\{\log(1-p)\}, & \text{for } \xi = 0, \end{cases} \quad (3.3.3)$$

where z_p denotes return level with return period of $1/p$. z_p is typically exceeded once every $1/p$ periods of time. On average, a rare event exceeds the z_p once every $1/p$ period. Using the GEVD, we can model the extremes of block maxima and minima. However, GEVD has the disadvantage of being unable to handle all the extreme data. Too small a block size can lead to poor asymptotic approximations.

Generalised Pareto distribution

EVT does not only focus on forecasting maxima or minima. It also focuses on forecasting the observations above a high threshold (Coles, 2001). Forecasting above a high threshold is useful when there is more than one large

value in a block and prevents ignoring some important values. The advantage of generalised Pareto distribution (GPD) is that it is more efficient than GEVD since it uses all the data (Singh et al., 2013). The GPD is also known as a POT distribution, and POT above the threshold μ is given by:

$$H_{\xi}(x) = \begin{cases} 1 - \exp\left(- \frac{\xi(x-\mu)}{\sigma} \right)^{-\frac{1}{\xi}}, & \xi = 0, \\ 1 - \left(1 + \frac{\xi(x-\mu)}{\sigma} \right)^{-\frac{1}{\xi}}, & \xi \neq 0, \end{cases} \tag{3.3.4}$$

where $x > \mu$, $y_+ = max(y, 0)$ and $\sigma_\mu > 0$, $\xi \in \mathbb{R}$. As a result, if the limit is real, excesses must occur in the GPD's field of attraction. The probability of unconditional survival is then calculated as follows:

$$Pr(X > x) = \phi_\mu[1 - Pr(X < x|X > \mu)], \tag{3.3.5}$$

where ϕ_μ denotes the probability of being above the threshold μ. The GEVD and GPD models are related, and it is worth demonstrating this. Suppose Y follows GEVD over a high threshold μ and y represents the excess $(x - \mu)$. If G is the distribution function assigned to Y in (3.3.1):

$$\begin{aligned} Pr(Y > \mu + y|Y > \mu) &= \frac{1 - G(\mu + y)}{1 - G(\mu)} \\ &= \frac{[1 + \xi(\mu - y + u)/\sigma]_+^{-1/\xi}}{[1 + \xi(\mu - u)/\sigma]_+^{-1/\xi}} \quad \text{as} \quad \mu \to \infty \\ &= [1 + \xi y/\sigma_\mu]_+^{-1/\xi} \end{aligned}$$

where $\sigma_\mu = \sigma + \xi(\mu - u)$. An appropriate threshold must be estimated when modelling extreme values using GPD. The GDP has the advantage of being able to handle all the extreme data. In general, it is not easy to estimate the two parameters of the GPD, the shape and scale parameters.

3.3.2 Extremal mixture model

Various extremal mixture models exist, such as parametric and non-parametric models. However, this study focuses on semi-parametric extremal mixtures. An extreme value mixture model can be classified into two groups: a model that describes data below the unextreme threshold and a model that describes the population distribution tail above the threshold. A bulk model below the threshold and a GPD model above the threshold are combined to form the mixed model. It is beneficial to use extreme value mixture models because they simultaneously capture both bulk distributions below the threshold and tail distributions above the threshold. Without wasting any information, it analyses all the available information. Choosing a flexible bulk model and tail model that simultaneously fits both the non-extreme and extreme data is one of the major aims of extreme value mixture models (Hu, 2013).

For the tail distributions, Behrens et al. (2004) developed a parametric model to fit the bulk distribution and a GPD to fit the extreme values. In estimating the threshold, the two distributions are spliced together. Many parametric distributions can be used as the bulk model, such as gamma, Weibull or normal. For observations below the threshold, the model can be extended to a mixture distribution by considering other parametric, semiparametric and non-parametric forms of distribution. This is the most straightforward among the extreme value mixture models. Parameter estimation can be carried out using all available data, including the threshold, which serves as a parameter. A model of this type can be especially useful since it incorpo-

rates threshold estimation and the corresponding uncertainty in the inference (Behrens et al., 2004).

In a study by Scarrott and MacDonald (2012), extremal mixture models for estimating sufficiently high thresholds are proposed to overcome the drawback of fixed threshold models, which do not capture threshold uncertainty. In their study, Scarrott and MacDonald (2012) grouped the threshold methods into the re-sampling-based, classical fixed threshold, tail fraction estimation and extremal mixture models. MacDonald et al. (2011) developed an extremal mixture model which integrates a non-parametric kernel density for the bulk distribution and GDP for the tail distribution. Uncertainty around parameter estimation and threshold selection is addressed by Bayesian inference (Sigauke and Bere, 2017). In another study, Fukutome et al. (2015) discussed an automatic threshold selection method. Sigauke and Bere (2017) developed a time-varying threshold coupled with a positive shift factor.

Semi-parametric extremal mixture model

The proposed modelling framework is based on the work of (Durrieu et al., 2015). Let $F_t(x) = P(X \leq x \mid T = t)$, where X is a random variable given at time index $t \in [0, T_{max}]$. The excess distribution function over a sufficiently high threshold μ is then given in Equation (3.3.6) (Durrieu et al., 2015, 2016):

$$F_{t,\mu}(x) = 1 - \frac{1 - F_t(x)}{1 - F_t(\mu)}, \qquad x \in [\mu, \infty), \tag{3.3.6}$$

where the cumulative distribution function (CDF) F_t is assumed to be in the domain of attraction of the Fréchet class of distributions. The CDF $F_{t,\mu}(x)$ can be approximated by a Pareto distribution, which is defined by Equation

(3.3.7).

$$G_{\mu,\theta}(x) = 1 - \left(\frac{x}{\mu}\right)^{-\frac{1}{\theta}}, \qquad x \in [\mu, \infty), \tag{3.3.7}$$

where the parameter $\theta > 0$ denotes the conditional tail index (extremal index) and an unknown threshold $\mu \geq x_0$, which depends on t. The empirical distribution function can approximate the distribution function F_t. The semi-parametric extremal mixture model (SPEM) is given in Equation (3.3.8) (Durrieu et al., 2015, 2016).

$$y_{t,\mu,\theta} = \begin{cases} 1 - \big(1 - F_t(x_t)\big)\big(1 - G_{\mu,\theta}(x_t)\big), & \text{if } x > \mu, \\ \\ F_t(x_t), & \text{if } x \in [x_0, \mu], \end{cases} \tag{3.3.8}$$

where $y_{t,\mu,\theta}$ is the global horizontal irradiance (GHI). From equation (3.3.8), $F_t(x_t)$ is a bulk model in which we fit a non-parametric model (kernel density) and $1 - \big(1 - F_t(x_t)\big)\big(1 - G_{\mu,\theta}(x_t)\big)$ is the tail model, in which we fit the Pareto distribution. We use the kernel estimator in the study to estimate the parameter θ. A point-wise data-driven procedure is used to choose a threshold μ. The bandwidth denoted by h will be determined by minimising the cross-validation function. Let $K(.)$ denote a kernel function assumed to be non-negative continuous and symmetric to the real line such that $K(x) \leq 1$. The set of weights for $i = 1, ..., m$ defined by the kernel function is given by Equation (3.3.9).

$$W_{t,h}(t_i) = K\left(\frac{t_i - t}{h}\right), \tag{3.3.9}$$

where bandwidth parameter $h > 0$.

Strengths

The SPEM discussed here has the advantage that it is based on one covariate, which is $t = 1, 2, ..., n$. It is capable of processing inputs of any length. The model size does not expand as the input grows larger. It is possible to share weights across any time step.

Limitations

On the other hand, SPEM has limitations on stability and accuracy. The quantity and position of the measured points have a significant impact on it. The size does not expand due to its recurrence. It is computationally expensive.

Based on semi-parametric extreme value mixture models

In this present study, the threshold selection is made by testing a goodness-of-fit test discussed by reference Durrieu et al. (2016) for the parametric part of the model in Equation (3.3.8). The k upper statistics are used to test the tail adjustment of the Pareto distribution. If the test is not rejected, the number k of upper statistics increases. Until the tail adjustment is rejected, it is tested again.

Based on non-parametric extreme value mixture models

A non-parametric density $H(.|X, \gamma)$, which depends on a parameter γ and the observation vector $\mathbf{X}$, is expected to be followed by the observations below the threshold. The upper tail is expected to follow a GPD or the point process representation outlined above. GPD and Non-parametric components are assumed to approximate the distribution of data generation (MacDonald

et al., 2011). The standard Kernel GPD model distribution function is given as:

Bulk model based tail fraction approach:

The cumulative distribution function takes the form (MacDonald et al., 2011):

$$F(x|X,\gamma,u,\sigma_u,\xi,\vartheta_u) = \begin{cases} H(x|X,\gamma) & x \leq u, \\ (1-\vartheta_u) + \vartheta_u \times \ddot{G}(x|u,\sigma_u,\xi) & x > u, \end{cases} \tag{3.3.10}$$

where $\vartheta_u = 1 - H(u|X,\gamma)$, $H(x|X,\gamma)$ denotes the kernel density estimator's distribution function with parameter γ. The GPD's distribution function signified by $\ddot{G}(x|u,\sigma_u,\xi)$ and ϑ_u is the bulk model-based tail fraction. The u, σ_u, and ξ denote the threshold, shape and scale parameters, respectively.

Parameterised tail fraction approach:

The purpose of the model (3.3.11) developed by MacDonald et al. (2011) is to provide extreme value analysis with a more flexible framework. The distribution function is written as follows:

$$F(x|X,\gamma,u,\sigma_u,\xi,\vartheta_u) = \begin{cases} (1-\vartheta_u)\frac{H(x|X,\gamma)}{H(u|X,\gamma)} & x \leq u, \\ (1-\vartheta_u) + \vartheta_u \times \ddot{G}(x|u,\sigma_u,\xi) & x > u, \end{cases} \tag{3.3.11}$$

Inference

The maximisation of the weighted quasi-log-likelihood estimator is defined by (Durrieu et al., 2015, 2016):

$$\hat{\theta}_{t,h,\mu} = \frac{1}{\hat{n}_{t,h,\mu}} \sum_{i=1}^{n} W_{t,h}(t_i) I_{\{X_{ti} > \mu\}} \log\left(\frac{X_{ti}}{\mu}\right), \tag{3.3.12}$$

where $\hat{n}_{t,h,\mu} = \sum_{i=1}^{n} W_{t,h}(t_i) I_{\{X_{ti}>\mu\}}$ represents the weighted number of observations above the threshold μ. The $F_t(x_t)$ is estimated by:

$$\hat{y}_{t,h,\mu} = \begin{cases} 1 - \big(1 - \hat{F}_{t,h}(\mu)\big)\big(1 - G_{\mu,\hat{\theta},h,\mu}(x_t)\big), & \text{if } x > \mu, \\ \hat{F}_{t,h}(x_t), & \text{if } x \in [x_0, \mu], \end{cases} \tag{3.3.13}$$

For any $\tau \in [0.99, 1)$ the estimator of the τ quantile of X_t is given by:

$$\hat{q}_\tau(t,h) = \begin{cases} \mu\big(\frac{1-\hat{\tau}_\mu}{1-\tau}\big)^{\hat{\theta}_{t,h,\mu}}, & \text{otherwise}, \\ \hat{F}^{-1}_{t,h}(\tau), & \text{if } \tau < \hat{\tau}_\mu, , \end{cases} \tag{3.3.14}$$

and $\hat{\tau}_\mu = \hat{F}_{t,h}(\mu)$.

3.3.3 Generalised additive extreme value model

The generalised additive extreme value (GAEV) model is a hybrid model which fits QR via the asymmetric Laplace distribution (ALD). The variation in extreme value distributions is represented here by a GAM framework. QR model values exceed a high threshold, corresponding to a fixed quantile. With GPD, subsequent excesses are modelled. Models such as GPD and Poisson-GPD require a high threshold. GAEV combines three things: flexibility in fitting extreme value distributions with GAM form parameters,

accurate inference for all functions and parameters for drawing common inferences from extreme value analyses, such as return level estimates based on quantified uncertainty (Youngman, 2019).

The high threshold $\mu(x)$ is estimated using QR based on a pre-specified quantile τ. The likelihood-based inference is achieved by comparing the tilted loss function used in QR and the asymmetric Laplace distribution (ALD). Let $Y(x)$ be a random variable indexed by a covariate x (Youngman, 2019, 2020). Then:

$$Y(x) \sim ALD\big(\mu(x), \sigma(x), \tau(x)\big), \tag{3.3.15}$$

Furthermore, the density function of ALD is given by (Youngman, 2019, 2020):

$$f_{ALD,\tau}\big(y(x), \mu(x), \sigma(x), \tau(x)\big) = \frac{\tau(x)(1-\tau(x))}{\sigma(x)} \exp\left\{-\rho_\tau(x)\left(\frac{y(x)-\mu(x)}{\sigma(x)}\right)\right\}, \tag{3.3.16}$$

where $\mu(x)$ corresponds to quantile $0 < \tau < 1$ at location x, $\sigma > 0$ and $\rho_\tau(y) = y(x)\big(\tau(x) - I\{y(x) < 0\}\big)$ denotes the check function, for indicator function I. Assume GAM forms in covariate x defined by (Youngman, 2019, 2020):

$$y_t = \beta_0 + \sum_{i=1}^{I}\sum_{j=1}^{J_i} \beta_{ij} b_{ij}(x_t) + \varepsilon_t, \tag{3.3.17}$$

where y_t is a GHI, β_{ij} denotes basis coefficients and b_{ij} is the basis functions and J_i denote the dimension of the spline basis to model the effect y_t.

Strengths

GAEV is a hybrid model which fits QR via ALD. It assumes GAM forms in covariate $X_t, t = 1, ..., n$. It is robust to extremely high quantiles and probabilities of rare events.

Limitations

The GAEV model requires a smoothing parameter. It has limitations on accuracy and stability. Extremely sensitive to the quantity and location of the measured points.

Estimation of the threshold

The log-likelihood for the quantile regression threshold model from Youngman (2019) is used to estimate the threshold. Log-likelihood based on fixed τ is defined by (Youngman, 2019):

$$\log f(y;\beta) = \sum_{j=1}^{J}\sum_{t=1}^{T} \log f\big(y(s_j,t); u(s_j,t), \sigma(s_j,t)\big), \tag{3.3.18}$$

where $\mathbf{u} = \big(u(s_1,1), ..., u(s_J,T)\big)' = \mathbf{X}_u\beta_u, \sigma = \big(\sigma(s_1,1), ..., \sigma(s_J,T)\big)', \log\sigma = \mathbf{X}_\sigma\beta_\sigma$, $\mathbf{X}_u$ and $\mathbf{X}_\sigma$ denote design matrices. $\mathbf{S}_{u,\lambda}$ and $\mathbf{S}_{\sigma,\lambda}$ are penalty matrices.

Inference: Restricted maximum likelihood

In order to make inference for β, penalised log-likelihood function is defined by (Youngman, 2019, 2020):

$$\ell(\beta_\lambda, \lambda) = \ell(\beta_\lambda) - \frac{1}{2}\beta' \mathrm{S}_\lambda \beta, \tag{3.3.19}$$

where $\lambda = (\lambda_1, ..., \lambda_I)$ represents the smoothing parameters, S_λ denotes a penalty matrix with elements determined by the basis function b_{ij} that has been chosen. And penalty matrix is defined by $S_\lambda = \sum_{i=1}^{I} \lambda_i S_i$, where S_i denotes rows and columns of the matrix corresponds to $b_{i'i}$, $i^{'} \neq i$ and consists of zeros. The parameters of extreme value distribution are estimated by a restricted maximum likelihood defined by (Youngman, 2019, 2020):

$$\ell(\beta_\lambda) = \ell(\hat{\beta}_\lambda, \lambda) + \frac{1}{2}\log|\mathrm{S}_\lambda|_+ - \frac{1}{2}\log|\mathrm{H}(\beta_\lambda^{'})| + \text{constant}, \tag{3.3.20}$$

where for given λ, $\hat{\beta}_\lambda$ miximises $\ell(\beta_\lambda, \lambda)$, $\mathrm{H}(\beta_\lambda^{'}) = -\nabla\nabla^{'}\ell(\beta, \lambda)|_{\beta=\hat{\beta}_\lambda}$ and $|\mathrm{S}_\lambda|_+$ is the product of positive eigenvalues of matrix S_λ.

Extremal index

The censored log-likelihood model is used for the estimation of an extremal index, which is given by (Youngman, 2019):

$$\log f(y;\beta_\theta) = \sum_{j=1}^{J}\left[\sum_{t=1}^{T} I\big(\tilde{w}(s_j,t) > w_u\big)\big[\log\{2\theta(s_j,t)\} - 2\log\{\tilde{w}(s_j,t)\}\big] \frac{2\theta(s_j,t)}{\max\{\tilde{w}(s_j,t), w_u\}}\right],$$

where $w_u = \frac{-1}{\log\tau}$.

3.3.4 Percentiles method

We first fit a penalised cubic regression smoothing spline given by (Sigauke and Bere, 2017):

$$\pi_1(x) = \sum_{t=1}^{n}\big(y_t - f(x_t)\big)^2 + \lambda\int\big(f^{''}(x)\big)^2 dx + \varepsilon_t, \tag{3.3.21}$$

where y_t denotes the response variable, λ represents a smoothing parameter and ε_t denotes the error term. We then extract residuals $\varepsilon_t = \pi_1(x) - \hat{\pi}_1(x)$ and then estimate the shift factor $\mu_\tau \in \Re,\ \tau \in [0.95; 1)$. The shift factor has to be sufficiently large to satisfy the asymptotic conditions when we fit the GPD (Sigauke and Bere, 2017).

$$\pi_2(x) = \sum_{t=1}^{n} \big(y_t - f(x_t)\big)^2 + \lambda \int \big(f''(x)\big)^2 dx + \varepsilon_t + \mu_\tau, \tag{3.3.22}$$

To estimate the shift factor μ_τ, extremal mixture models will be used.

Strengths

A large sample approximation is used for inference. The response variable is resilient to outliers. The transparent process makes it easy to implement and debug. No training period; it is called a lazy learner. It is simple to add new data.

Limitations

A smoothing parameter is necessary. Only the extreme quantiles are impacted by outliers or the quantile curves close to them. Fails miserably with a big dataset. Features must be scaled. Sensitive to missing numbers, noisy data and outliers.

3.3.5 Combining the extreme quantiles

This study uses simple average and median methods to combine the extreme quantiles from the proposed models. The $100(1-\alpha)\%$ combined forecasts interval for $\tilde{x}$ developed from the k individual intervals with heuristic H is denoted by $[L_H, U_H]$ (Gaba et al., 2017).

Simple average method

A simple average method performs better than other sophisticated methods when combining forecasts. The endpoints of the intervals are averaged using this procedure. Simple averages are used in combining point forecasts due to their robustness, simplicity and providing accurate forecasts. The lower and upper averages models are defined by $L_{AV} = (1/k)\sum_{i=1}^{k} L_i$ and $U_{AV} = (1/k)\sum_{i=1}^{k} U_i$ respectively (Gaba et al., 2017).

Median method

The median method is known for summarising data because, compared to the mean, it is less susceptible to extreme values. The median models are given by $L_{Md} = Median\{L_1, ..., L_k\}$ and $U_{Md} = Median\{U_1, ..., U_k\}$ (Gaba et al., 2017).

3.4 Multivariate Adaptive Regression Spline

The multivariate adaptive regression spline (MARS) model is a non-parametric multivariate regression method developed by Friedman (1991), which allows a flexible regression analysis for high-dimensional data. The model is based on an expansion of product spline basis functions. The data automatically determines the parameters (product degree and knot locations) and the number of basis functions. Compared to other techniques, such as artificial neural networks, classification and regression trees, the MARS model has better interpretability characteristics and is easy to understand. Modelling space in MARS is divided into subregions, and then simple linear regression models are fitted in each subregion. A model is built in the forward and backward

stepwise algorithms. MARS constructs many basis functions in the stepwise forward step, which overfits the data. Generalised cross-validation (GCV) is used in the backward stepwise step to delete the basis functions based on least contribution (Craven and Wahba, 1979). The general MARS model is defined by (Friedman, 1991):

$$f(x) = \beta_0 + \sum_{m=1}^{M} \beta_m B_m(x), \tag{3.4.1}$$

where

$$B_m(x) = \prod_{k=1}^{K_m} \left[s_{km} \left(x_{v(k,m)} - t_{km} \right) \right],$$

denotes a basis function, β_0 and β_m are parameters, M represents the number of basis functions, K_m represents the number of knots, s_{km} is either 1 or -1, depending on the associated step function of right or left sense, $v(k,m)$ denotes the label of the independent variable and t_{km} denotes the knot location. The GCV criterion considers residual error and model complexity as a measure of goodness of fit. The GCV criterion is given by (Craven and Wahba, 1979):

$$\text{GCV}(M) = \frac{\frac{1}{N} \sum_{i=1}^{N} \left[y_i - \hat{f}_M(x_i) \right]^2}{\left[1 - \frac{C(M)}{N} \right]^2}, \tag{3.4.2}$$

where N represents the sample size, $C(M)$, is a measure of the cost-penalty associated with a model containing M basis functions, $\hat{f}_M(x_i)$ measures the lack of fit on the basis function model of M and $C(M)$ is the penalty for model complexity in the denominator. The complexity cost function is defined by (Friedman, 1991):

$$C(M) = \text{trace}\left(B(B^T B)^{-1} B^T \right) + 1, \tag{3.4.3}$$

where B denotes the M $\times$ N data matrix of the M (non constant) basis functions ($B_{ij} = B_i(x_j)$). The model with the lowest GCV criterion value is the best.

3.5 Conditional Multivariate Extreme Value Modelling

Copulas, joint tail models and other multivariate extreme approaches rely on the assumption that all variables will be enormous simultaneously. The conditional extremes model (CEM) overcomes this limitation (Heffernan and Tawn, 2004; Keef et al., 2013; Tilloy et al., 2020). Using a CEM, the dependence structure among variables is estimated where one variable is conditioned on being extreme and attempts to model the conditional distribution (Heffernan and Tawn, 2004; Keef et al., 2013). This study uses the conditional multivariate approach (Heffernan and Tawn, 2004).

3.5.1 Threshold Selection

In GPD, we are interested in the observations above a sufficiently high threshold. The threshold selection is comparable to the block size selection in the block maxima method, and bias and variation must be balanced. (Coles, 2001). This study will use extremal mixture models. The benefit of utilising extreme value mixing models is that they simultaneously capture the bulk distribution below the threshold and the tail distribution above the threshold (Scarrott and MacDonald, 2012; Hu, 2013).

3.5.2 Bivariate threshold excess model

EVT focuses not only on modelling maxima or minima data. It also focuses on predicting the observations which are above a high threshold. The multivariate modelling in this study is limited to pairwise combinations of variables. GPD is a model used in EVT for exceedances above a threshold and is defined by equation (3.3.4). Where $\eta = \Pr(X > u)$ for a family represented on $\left[1 + \xi\left(\frac{x-u}{\sigma_u}\right)\right] > 0$ and $x - u > 0$. Therefore, $F(x) \approx G(x)$ on $x > u$ for a sufficiently high threshold u with parameters η, ξ and σ_u (Coles, 2001). The objective is to derive a family that approximates a joint distribution $F(x, y)$ on regions $x > u_{x,y} > u_y$ for sufficient large thresholds u_x and u_y.

3.5.3 Marginal transformation: Laplace margins

The margins are first transformed into standardised Laplace margins in the regression-type dependence model before dependence models are built. There are symmetric and exponential tails in the Laplace distribution. Laplace margins simplify the regression structure compared to other transformations, such as Fréchet and Gumbel margins (Heffernan and Tawn, 2004; Keef et al., 2013). Therefore, both positive and negative dependence can be described by a single model structure (Maposa et al., 2021). Let

$$Y_i = \begin{cases} \log\left\{2F_i(X_i)\right\}, & \text{for } X_i < F_i^{-1}(0.5) \\ -\log\left\{2\left[1 - F_i(X_i)\right]\right\}, & \text{for } X_i > F_i^{-1}(0.5) \end{cases} \tag{3.5.1}$$

for $i \in D = \{1, 2, \ldots, d\}$. The vector $\mathbf{Y} = (Y_1, \ldots, Y_d)$ is then said to have a Laplace distribution given by (Keef et al., 2013):

$$P(Y_i < y) = \begin{cases} e^{\frac{y}{2}}, & \text{if } y < 0 \\ 1 - e^{-\frac{y}{2}}, & \text{if } y \geq 0, \end{cases} \tag{3.5.2}$$

for all $i \in D$.

3.6 Bivariate Dependence Modelling Using Copulas

A copula approach allows us to model different types of relationships across a wide range of behaviour. In analysing the dependence structure, the copula function renders many advantages. First, copulas allow the modelling of marginal behaviour and dependency structures separately. Second, the copula function can show the dependence's degree and structure. This allows for capturing asymmetric dependency since a linear correlation does not reveal information regarding tail dependence. Thirdly, copulas do not imply elliptically distributed random variables, unlike correlation (Naifar, 2011). There are various copulas, such as the Archimedean, extreme value and elliptical copulas. This study will use Archimedean copulas due to their advantage in capturing dependencies in the upper and lower tails.

3.6.1 Sklar's Theorem

Sklar's theorem is considered the central theorem of copula theory, whereby the n-dimensional distribution function can be divided into two basic components, the marginal distribution and the copula (Nelson, 2006).

Sklar's theorem: Let H denote an n-dimensional distribution function with marginals F_1, F_2, ..., F_n. Then for all $x_1, x_2, \ldots, x_n \in \mathbb{R}$, there is an n-

copula C,

$$H(x_1, x_2, \ldots, x_n) = C\big(F_1(x_1), F_2(x_2), \ldots, F_n(x_n)\big). \tag{3.6.1}$$

The corresponding theorem in two dimensions is derived for $n = 2$. This important theorem shows that copula functions are essential for multivariate distribution functions whose marginals are either known or given. Based on Sklar's theorem, there is a copula C:$[0, 1]^n \rightarrow [0, 1]$. It maps the multivariate marginal distribution F_i to the univariate distribution function F. The multivariate joint density function is defined by (Mupondo et al., 2021):

$$f(x_1, \ldots, x_n) = c(F_1(x_1), \ldots, F_n(x_n)) \times f_1(x_1) \times \ldots \times f_n(x_n). \tag{3.6.2}$$

3.6.2 Archimedean Copulas

Archimedean copulas are a special type of copulas. Their properties have enabled them to be used in various areas because they are easy to construct (Nelson, 2006). Compared to traditional techniques, Archimedean copulas are designed to produce a significantly better dependency model because they are more tractable mathematically. In addition, they can construct multivariate copulas using a simple hierarchical method, which is often ineffective with other copula classes (Corbella and Stretch, 2013). Archimedean copulas contain dependence models that are viable for modelling upper and lower tail dependences. The Archimedean copula equation is given by (Naifar, 2011):

$$C_{arch}(u_1, u_2) = \Phi^{-1}\big(\Phi(u_1) + \Phi(u_2)\big), \tag{3.6.3}$$

where Φ denotes a generator function of the copula, if for all $0 \leq u_1, u_2 \leq 1$. Numerous families present different dependency structures within the

Archimedean copula, simplifying the construction of bivariate distributions (Naifar, 2011). This study employs three Archimedean copula functions: Clayton, Frank and Gumbel.

The Clayton copula

Clayton copula was proposed by Clayton (1978) and is given in Equation (3.6.4).

$$C_{\theta}^{Clayton}(u_1, u_2) = (u_1^{-\theta} + u_2^{-\theta} - 1)^{-\frac{1}{\theta}}, \tag{3.6.4}$$

where $\theta \in [-1, \infty)/\{0\}$ controls the degree of dependence between u_1 and u_2. The bivariate Clayton copula function converges to the comonotonic copula if $\theta = 0$ and tail independence if $\theta \to 0$. If $\theta = -1$, we obtain the lower bound of Fréchet-Hoeffding. Kendall's tau is defined as the bivariate Clayton copula function parameter as follows:

$$\tau_k = \frac{\theta}{\theta + 2}. \tag{3.6.5}$$

The lower tail dependence is estimated using the following function:

$$\lambda^L = 2^{-\theta}. \tag{3.6.6}$$

The Frank copula

Frank copula was proposed by Frank (1979) and is given in Equation (3.6.7).

$$C_{\theta}^{Frank}(u_1, u_2) = -\frac{1}{\theta} \ln \left(1 + \frac{(e^{-\theta u_1} - 1)(e^{-\theta u_2} - 1)}{(e^{-\theta} - 1)} \right), \tag{3.6.7}$$

where θ is any real number. Based on the dependence parameter (θ), the upper and lower Fréchet-Hoeffding bounds can be estimated. The model is not tail dependent ($\lambda_{u_1} = \lambda_L = 0$) and is suitable for modelling data with weak tail dependency.

The Gumbel copula

Gumbel copula was proposed by Gumbel (1960) and is given in Equation (3.6.8).

$$C_{\theta}^{Gumbel}(u_1, u_2) = \exp\left(-\left[(-\ln u_1)^{\theta} + (-\ln u_2)^{\theta}\right]^{\frac{1}{\theta}}\right); \qquad 0 \leq u_1, u_2 \leq 1, \tag{3.6.8}$$

where $\theta \in [1, \infty)$ determines the degree of dependence between u_1 and u_2. When $\theta = 1$, the bivariate Gumbel copula converges to complete independence, and when $\theta \to 0$, perfect independence is achieved. Kendall's tau (τ) and the bivariate Gumbel parameter (θ) are linked by the following formula:

$$\tau_k = 1 + \theta^{-1}. \tag{3.6.9}$$

To estimate the upper (λ^U) and lower (λ^L) tail dependence of the bivariate Gumbel copula, the following functions are used:

$$\lambda^U = 1 + 2^{-\theta} \text{ and } \lambda^L = 0.$$

3.6.3 The bivariate Archimedean copula parameter (θ) and its relationship with the Kendall's tau (τ) and tail dependence (λ^U and λ^L)

As discussed in Naifar (2011), Archimedean copula functions are attractive due to their relationship with the tail dependence coefficient and Kendall's tau. Table 3.1 presents the relationship between bivariate Archimedean copula parameters (θ), tail dependence coefficients: Kendall's tau (τ) and tail dependence (λ^U and λ^L). The Clayton copula is an asymmetric copula that exhibits lower tail dependence. The Gumbel copula function represents the upper tail dependence, which is also an asymmetric copula. This indicates

that the compound variables are moving upwards rather than downwards at the same time. The tail dependence coefficient (λ^U and λ^L) is a measure of how dependent two variables are on each other's tails.

Table 3.1: The relationship between the bivariate Archimedean copula parameter (θ) and tail dependence coefficients: Kendall's tau (τ) and tail dependence (λ^U and λ^L).

Family	**Copula Parameter (Θ)**	**Kendal's Tau (τ)**	**Upper Tail (λ^U)**	**Lower Tail (λ^L)**
Clayton	$\theta > 1$	$\frac{\theta}{\theta+2}$	0	$2^{-\theta}$
Frank	$-\infty < \theta < \infty$	$1 - \frac{4}{\theta}[D_j(\theta)]$	0	0
Gumbel	$\theta \geq 1$	$\frac{\theta-1}{\theta}$	$2 - 2^{-\theta}$	0

3.6.4 Parameter estimation and goodness of fit tests

The Archimedean copula functions in this study have only one dependence parameter, which will be estimated using the maximum likelihood method. The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) will be used to select the best-fitting Archimedean copula models. The AIC and BIC are given by:

$$\text{AIC} = 2\ell + 2k, \tag{3.6.10}$$

and

$$\text{BIC} = 2\ell + \ln(n)k, \tag{3.6.11}$$

where ℓ represents the log-likelihood and ln the natural logarithm.

3.6.5 Mixture of copulas

It is more flexible to model various dependence structures using a mixture of copula functions than using individual copula functions. Mixed copulas are

defined by their association parameters, with weights on each copula function indicating the strength of dependence (Hu, 2006). Several efficient computational methods have allowed mixture models to be used for clustering in recent years. Copulas are also growing in popularity as a model for simulating dependency since they distinguish between the dependent attributes and the marginal properties of data (Arakelian and Karlis, 2014). Copulas are general measures of dependencies, and the parameters they contain can be used to estimate how our actions and markets affect the world around us. Mixed Copula provides several additional advantages when measuring random variable dependence. First, the mixed Copula can join any dependence structure because it is constructed from various Copula functions. Second, although the data is converted into different kinds, the dependent structures that mixed copulas capture remain the same (Yamaka et al., 2021).

Let $C_i(u_1, u_2)$ denote a copula $i,\ i = 1, \ldots, m$, i.e., m copulas of dimension d. A mixture of these $d-$dimensional copulas with weights $w_i,\ i = 1, \ldots, m$ is also a $d-$dimensional copula whose distribution function is given by Equation (3.6.12).

$$C^{\text{mix}}(u_1, u_2) = \sum_{i=1}^{m} w_i C_i(u_1, u_2), \tag{3.6.12}$$

with $\sum_{i=1}^{m} w_i = 1$, i.e., the mixture is a convex combination of the m copulas. From Equation (3.6.12) we have

$$\begin{aligned} C_{FG}^{\text{mix}}(u_1, u_2) &= \sum_{i=1}^{2} w_i C_i(u_1, u_2) \\ &= w_1 C_1(u_1, u_2) + w_2 C_2(u_1, u_2) \\ &= w_1 C_{\theta_1}^{\text{Frank}}(u_1, u_2) + w_2 C_{\theta_2}^{\text{Gumbel}}(u_1, u_2) \end{aligned}$$

Since $\sum_{i=1}^{m} = 1$, we have $w_1 + w_2 = 1 \implies w_2 = 1 - w_1$. This leads to

$$\begin{aligned} C_{FG}^{\text{mix}}(u_1, u_2) &= w C_{\theta_1}^{\text{Frank}}(u_1, u_2) + (1-w) C_{\theta_2}^{\text{Gumbel}}(u_1, u_2) \\ C_{FG}^{\text{mix}}(u_1, u_2) &= w \left[-\frac{1}{\theta_1} \ln \left(1 + \frac{(e^{-\theta_1 u_1} - 1)(e^{-\theta_1 u_2} - 1)}{(e^{-\theta_1} - 1)} \right) \right] + \\ & \quad (1-w) \left[e^{\left(-\left[(-\ln u_1)^{\theta_2} + (-\ln u_2)^{\theta_2} \right]^{\frac{1}{\theta_2}} \right)} \right] \end{aligned}$$

3.7 Error Measures for Probabilistic Forecasting and Evaluation of Methods

The approaches for analysing and contrasting probabilistic forecasts derived from extreme quantile models are presented in this section. The scoring rule assigns a penalty score represented by $S(y, F)$ to the probabilistic forecast, where y denotes the observation used for forecast assessment and F represents the forecast distribution (Ghofrani and Alolayan, 2018). When the quantile score is low, the forecasting model is more accurate. In this study, three error measures will be used, such as continuous rank probability score (CRPS), the Dawid-Sebastiani score (DSS) and the pinball loss function (PLF).

3.7.1 Continuous rank probability score

The CRPS measures the difference between the scalar variable's cumulative density functions that are expected and those that are observed (Ghofrani and Alolayan, 2018). The CRPS is defined by:

$$\mathrm{CRPS}(y, F) = \int_0^1 \mathrm{QS}_\tau \left(F^{-1}(\tau), y \right) d\tau, \tag{3.7.1}$$

where F represents the forecast distribution and QSτ denotes the quantile score defined by:

$$\mathrm{QS}_\tau \left(F^{-1}(\tau), y \right) = 2 \left(\mathbf{I} \left[y \leq F^{-1}(\tau) \right] \; - \; \tau \right) \left(F^{-1}(\tau) \; - \; y \right), \tag{3.7.2}$$

where $\mathbf{I}$ denotes an indicator function.

3.7.2 Dawid-Sebastiani score

The drawback of CRPS is that it is difficult to compute complex forecast distributions. The DDS is the alternative solution because it overcomes the drawback of CRPS by computing the complex forecast distribution easily. It is given by (Dawid and Sebastiani, 1999):

$$\mathrm{DSS}_{y,F} = \frac{(y \; - \; \mu_F)^2}{(\sigma_F^2)} + 2 \log(\sigma_F), \tag{3.7.3}$$

where F denotes forecast distribution, with the mean and standard deviation given by μ_F and σ_F^2, respectively.

3.7.3 Pinball loss function

Even if CRPS is useful for measuring the cumulative distribution function forecast quality, the study by Taillardat et al. (2019) indicated that the

CRPS is more applicable to distributions with smaller uncertainty intervals. However, a more suitable scoring rule is the PLF or quantile loss (Friederichs and Thorarinsdottir, 2012). The PLF is relatively easy to use and is defined by the following:

$$L(q_\tau, y_t) = \begin{cases} \tau(y_t - q_\tau), & \text{if } y_t > q_\tau, \\ (1-\tau)(q_\tau - y_t), & \text{if } y_t \leq q_\tau, \end{cases} \tag{3.7.4}$$

where y_t denotes the observed value of the response variable and q_τ represents quantile forecast.

3.8 Unit Commitment

Unit commitment (UC) minimises generating units' total cost within a specific time or interval (Wright, 2013). According to Aghaei et al. (2013), different methods and algorithms used to determine the optimal solution of the UC problems are classified into deterministic, heuristic, and hybrid approaches. This study will use mixed-integer linear programming (MILP), which falls under the deterministic approach. MILP is a special class of linear programming Wright (2013); Lebotsa et al. (2018), where variables are made of integer and continuous variables (Aghaei et al., 2013). This study envisions demonstrating how to use forecasts to solve the UC problem. Let P_{Gi}^{ht} be Northern Cape system load at hour h, h = 18:00; 19:00; 20:00; 21:00 on the day t, $t = 1, ..., n$, $P_{Gi}^{ht}(min)$ is the lower limit of the power output, $P_{Gi}^{ht}(max)$ represents the upper limit of the unit power output, x_i^{ht} be the 0 - 1 variable (This study assumes that during the peak period, all units are up, i.e. $x_i^{ht} = 1$ for all units), F_{si} is the start-up cost of unit i at hour h (for this

study we assume that the start-up cost is zero), P_R^{ht} is the power reserve at hour h on the day t and F_i is the average production cost of unit i (cost/MW). This study will use the fuel cost to represent the average production cost per megawatt.

3.8.1 Problem formulation using mixed-integer linear programming

Generating units are grouped into one hydroelectric denoted with a subscript d, eight wind power denoted with a subscript w, eight concentrated solar power (CSP) denoted with a subscript c and nineteen solar photovoltaic (PV) power denoted with a subscript p. These will be denoted as:

$g_{1d}; g_{2w}, g_{3w}, \ldots, g_{9w}; g_{10c}, g_{11c}, \ldots, g_{17c}; g_{18p}, g_{19p}, \ldots, g_{36p}.$

The objective function to minimise generating unit cost over a specific time is given by (Wright, 2013; Lebotsa et al., 2018):

$$\min \sum_{h=1}^{H} \sum_{i=1}^{m} \left[F_i \left(P_{Gi}^{ht} \right) x_i^{ht} + F_{si}(ht) x_i^{ht} \right] = F \left(P_{Gi}^{ht}, x_i^{ht} \right), \tag{3.8.1}$$

The constraints are defined as follows:

Load balance equation

$$\sum_{i=1}^{m} P_{Gi}^{ht} x_i^{ht} = P_D^{ht}, \quad h = 18,\ 19,\ 20,\ 21, t = 1, ..., n, \tag{3.8.2}$$

Generator power output limits

$$x_i^{ht} P_{Gi\min} \leq P_{Gi}^{ht} \leq x_i^{ht} P_{Gi\max},\ i = 1, ..., m \tag{3.8.3}$$

In this study total number of generating units, $m = 36$.

Chapter 4

Economic dispatch of electrical power in South Africa: An application to the Northern Cape province

4.1 Introduction

Utility companies rely on forecasting to operate their electricity supply systems (Arriagada et al., 2015). The objective of economic dispatch is to operate the power system with a minimum amount of operating costs. It also provides the important aspects of power system operation, such as meeting load demand at minimum cost by scheduling the committed generating units, reducing the emissions, maintaining the system stability, and security restriction (Al Farsi et al., 2015). Electricity load forecasting is important for economic dispatch because it provides future electricity production and consumption (Fan and Hyndman, 2012), which helps electricity utility to maintain the balance of demand and supply (Pierrot and Goude, 2011). Electricity load forecasting is also important for trading and production planning

in the electricity markets. It has different implementations, such as energy acquiring and production, load switching, contract rating and infrastructure evolution. Load forecasting helps electric utility management plan the distribution of electricity (Cavallaro, 2005; Mohandes, 2002; Singh et al., 2014).

Increasingly sophisticated technologies, including electric cars, smart grids and renewable energy production, pose challenges for electricity load forecasting. Modelling hourly extreme peak electricity demand at the high tail quantiles of the distribution is discussed in this study. Peak electricity is the highest load at a given time. Peak electricity load forecasting is essential because it assures the availability of enough supply. Under forecasting of peak electricity load results in an insufficient capacity for meeting blackouts and load demand. Power blackouts are a problem because they affect the operation of the economy. Extreme peak electricity load forecasting is the solution to the underprediction of the peak electricity load demand.

4.2 Empirical Results

Results from predicting the demand for electricity at 18:00, 19:00, 20:00 and 21:00 are shown in this study. The data used in this study is hourly electricity load demand for the Northern Cape, South Africa, from 01 January 2000 to 31 March 2014, giving a sample size of n = 5204, which represents the number of days. The predictor variables are temperature, non-linear trend, and day type. In contrast, the response variable is the load at hours 18:00, 19:00, 20:00 and 21:00. Hourly temperature data used is from De Aar weather station located at coordinates -30.6500, 24.0167, and its elevation is

1247 m and Calvinia weather station located at coordinates -31.500, 19.726, and its elevation is 990.6 m in Northern Cape, South Africa.

The temperature data were aggregated to get each station's maximum, minimum, and average daily temperature. The maxima, minima, and average daily temperature for both stations were then combined to get the average of the maxima (AveMaxT), average of the minima (AveMinT) and average daily temperature (AveTem) of the province that they represent. The response variable was fitted with a penalised cubic smoothing spline to determine the non-linear trend (nonlintrend18, nonlintrend19, nonlintrend20, and nonlintrend21). Day type variable denotes the days of the week, coded as 1 for Monday, 2 for Tuesday, to 7 for Sunday.

4.2.1 Exploratory data analysis

Table 4.1 presents the data for the hours of 18:00, 19:00, 20:00, and 21:00 for the sampling period from 1 January 2000 to 31 March 2014. The maximum demand (in mega watts-MW) is 880, 905, 884, and 855 for each of the four hours. It is confirmed that the distributions of the four hours are not normally distributed because the mean and median are not equal. The skewness and kurtosis further demonstrate the non-normality of the distributions.

Table 4.1: Summary statistics for hours 18:00, 19:00, 20:00, and 21:00 for electricity demand (MW).

Hour	Mean	Median	Min	Max	Std	Skewness	Kurtosis
18:00	676.7	684	397	880	83.15992	-0.2860717	-0.3457701
19:00	707.8	714	403	905	86.97296	-0.3078065	-0.3036559
20:00	715.9	727	434	884	74.94497	-0.4881771	-0.2331249
21:00	690.6	699	439	855	69.54004	-0.3849729	-0.3146267

Figure 4.1 displays the time series and density plots for the Northern Cape electricity demand at 18:00, 19:00, 20:00, and 21:00. The time series plots are shown in the left panels of Figure 4.1, while the demand density plots for hours under consideration are shown in the right panels. The left panels of Figure 4.1 depict the seasonal patterns of power demand in the Northern Cape, which show that demand is higher in winter and lower in summer annually. On four hour time series plots, there is a decrease in demand between 2008 and 2009. There are several possible reasons for the decrease in electricity demand. The use of renewable energies, energy efficiency appliances, and demand side management are some of the ones. Some big companies such as Sasol rely upon their own self generated electricity, which is used by the company for its own needs. The densities of the four hours on the right panels of Figure 4.1 indicate that there is non-normality in the distributions, which supports the report of kurtosis and skewness in Table 4.1. Figure 4.2 highlighted electricity demand for four hours in the box and whisker plots. Figure 4.3 shows the normal Q–Q plots of electricity demand for four hours. The plots show strong seasonality with upward trends. It indicate that the data is not normally distributed. Figure 4.4 presents the plots of electricity demand for four hours superimposed with a non-linear trend.

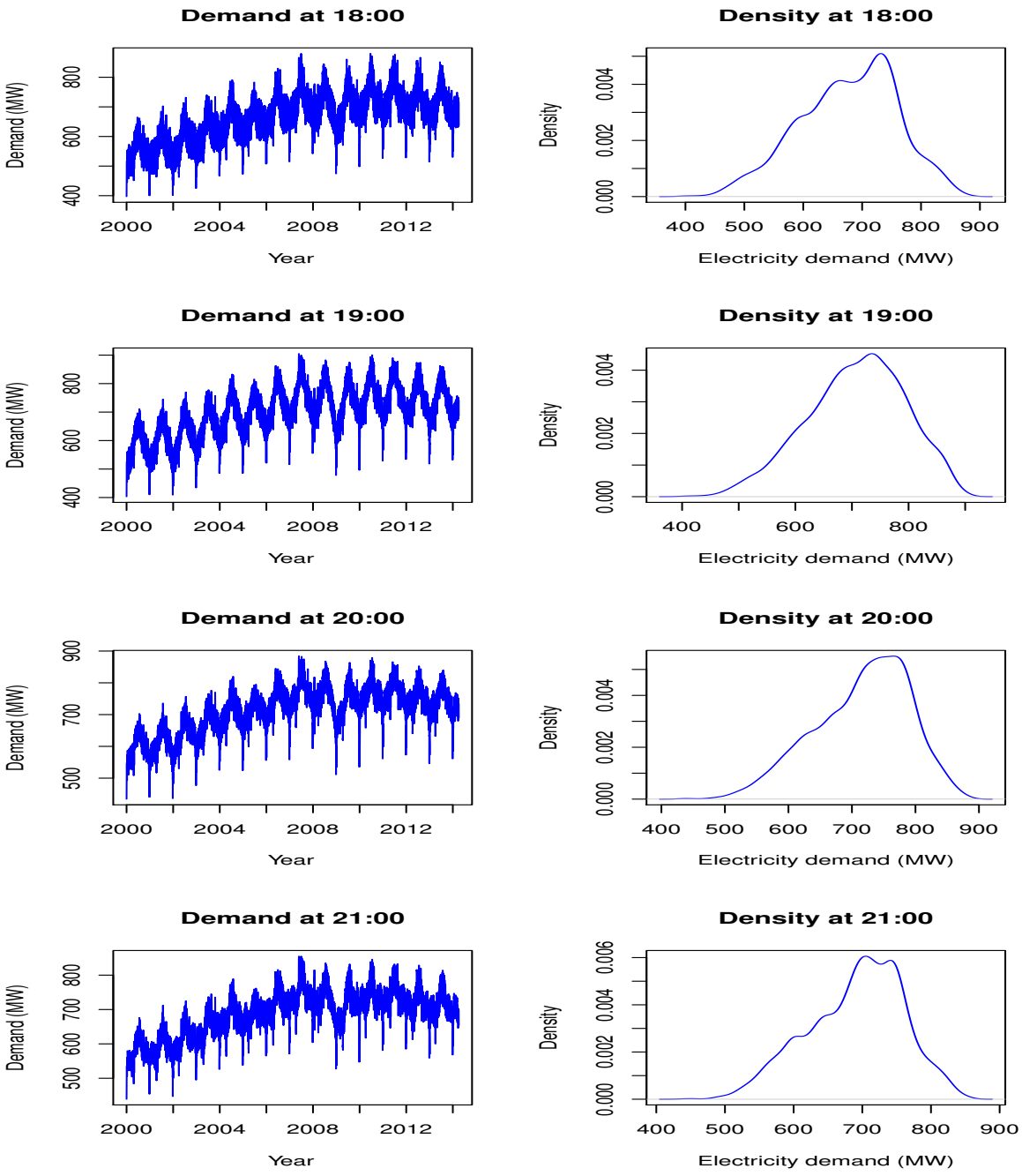

Figure 4.1: Plot of electricity demand and densities at 18:00, 19:00, 20:00, and 21:00.

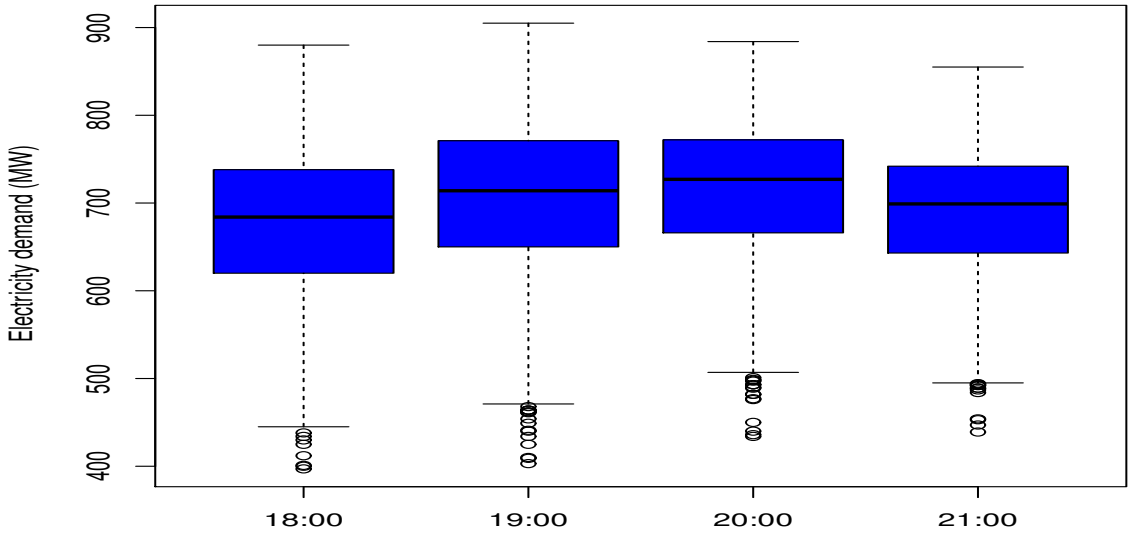

Figure 4.2: Box plots for hours 18:00, 19:00, 20:00, and 21:00.

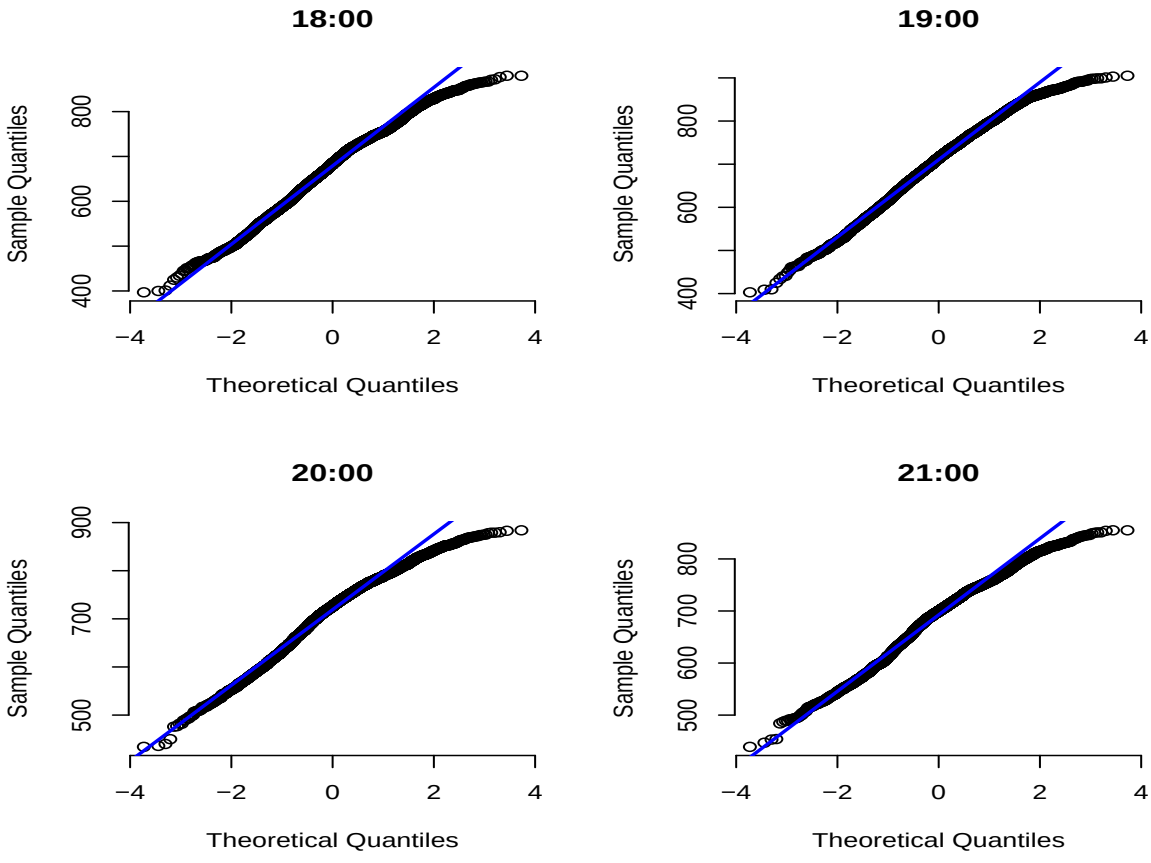

Figure 4.3: Normal Q–Q plots for hours 18:00, 19:00, 20:00, and 21:00.

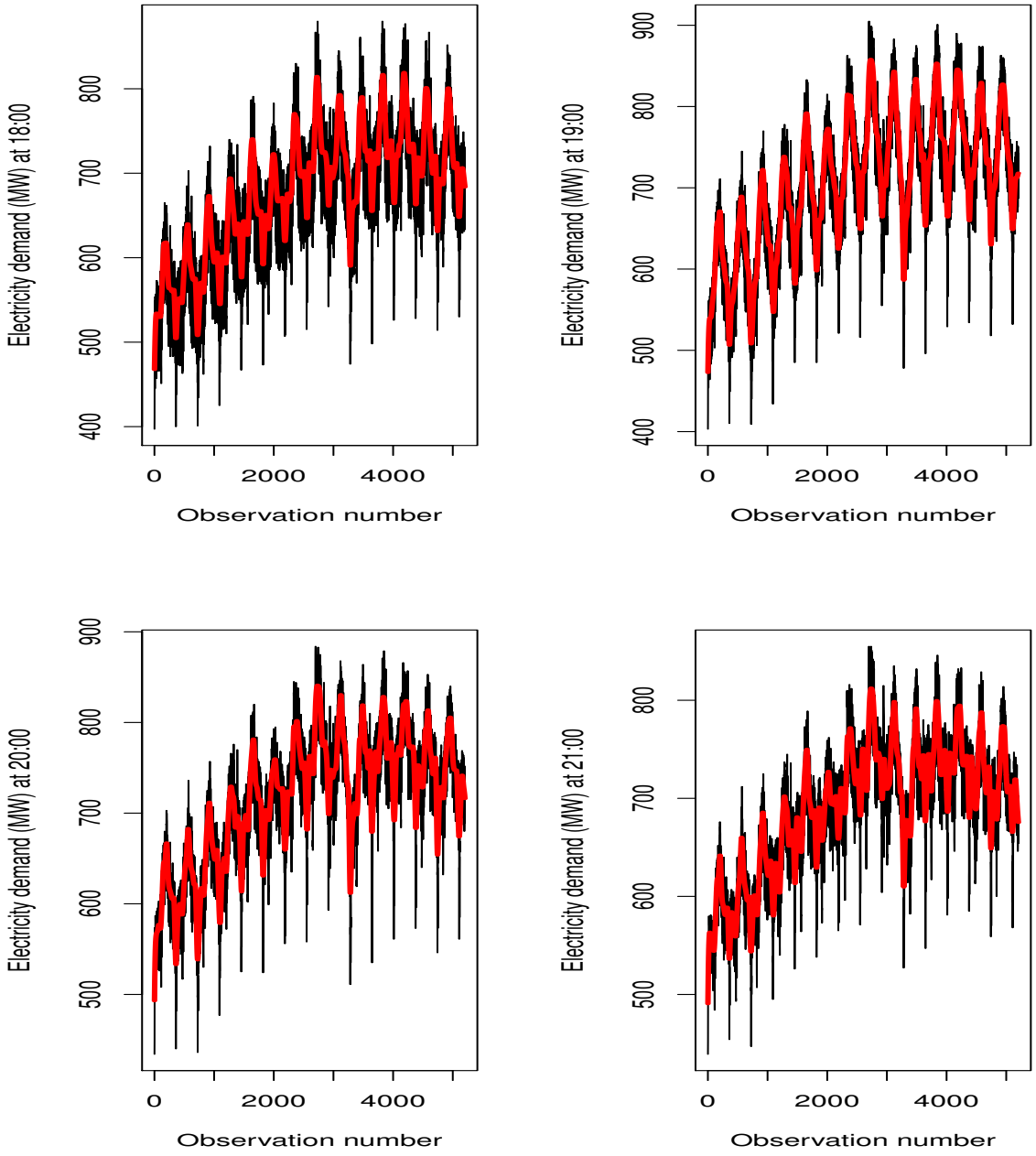

Figure 4.4: Plot of electricity demand at hours 18:00, 19:00, 20:00, and 21:00 with a non-linear trend (red curve).

4.2.2 Forecasting results

Forecasting electricity demand when covariates are given

The data are for 01 January 2000 to 31 March 2014, giving a sample size of n = 5204 observations. The data are split into two parts. The data for the period 01 January 2000 to 25 May 2011 are used for training, which is 80% of the data ($n_1 = 4163$), and the period from 26 May 2011 to 31 March 2014 is used for testing, which is 20% of the data ($n_2 = 1041$). The models considered are linear quantile regression (LQR), non-linear quantile regression (NLQR), and additive quantile regression (AQR). On the testing data, pinball loss was used to evaluate forecasts. Based on the pinball loss presented in Table 4.2, the best fitting model is AQR for hours 18:00 and 19:00 and LQR for hours 20:00 and 21:00. This is done for $\tau = 0.9999$ using R-package "qgam" for AQR, and "quantreg" for LQR, and NLQR.

Table 4.2: Summary table for Pinball loss for $\tau = 0.9999$

Model	18:00	19:00	20:00	21:00
LQR	0.014980	0.013776	**0.012544**	**0.01146**
NLQR	0.027425	0.020668	0.019867	0.020587
AQR	**0.013498**	**0.013441**	0.012827	0.011964

The best-fitting models were utilised to forecast the electricity demand after variable importance and variable selection using the least absolute shrinkage and selection operator (Lasso). The plots of actual demand and forecasts from the developed models are for hours 18:00, 19:00, 20:00, and 21:00 in Figure 4.5. It is demonstrated in Figure 4.5 that the forecasts from each model follow the actual demand data remarkably well at a high quantile.

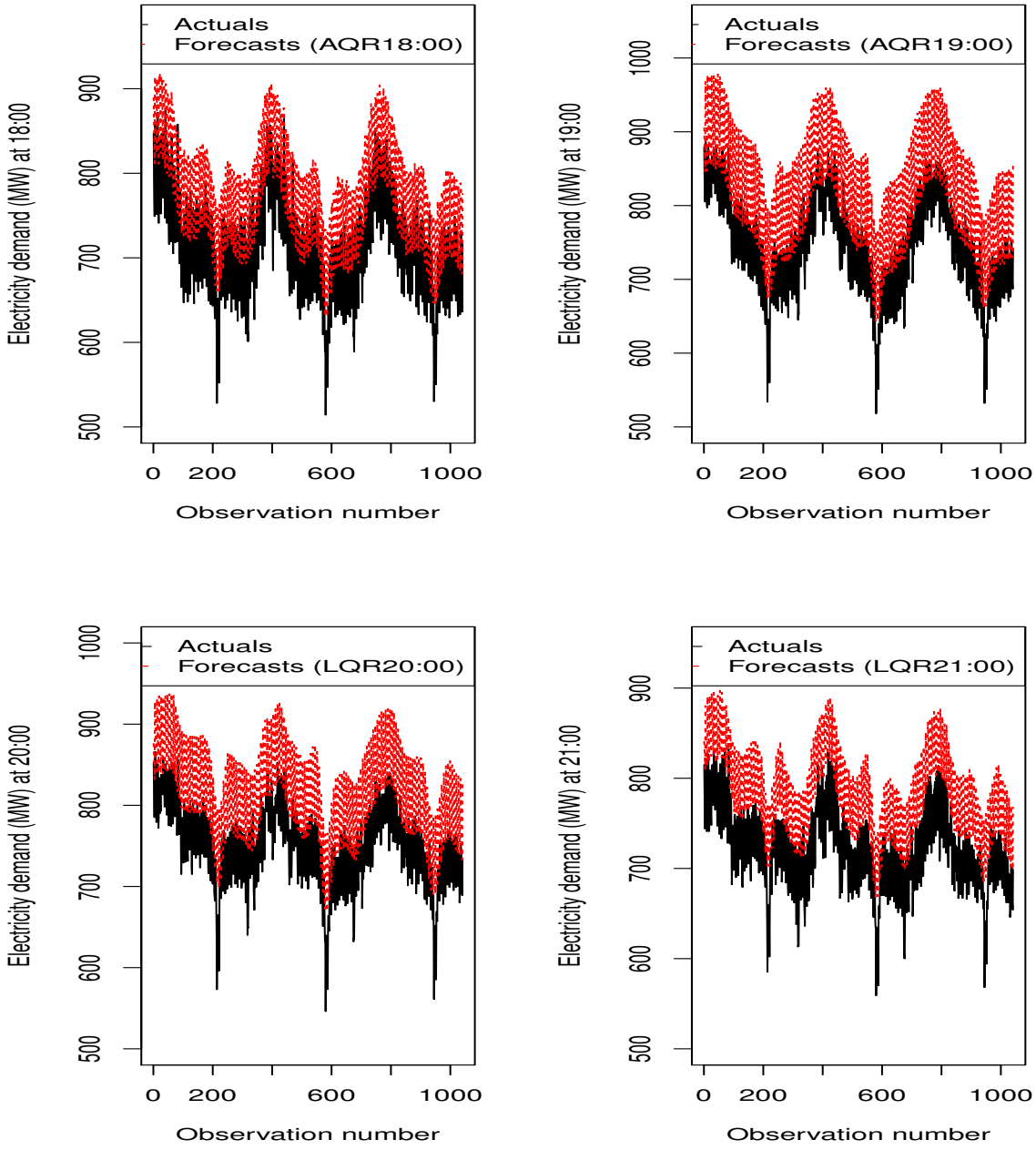

Figure 4.5: Plot of actual demand and forecasts at 18:00, 19:00, 20:00, and 21:00.

Forecasting electricity demand when covariates are not given

The predictor variables Daytype, Month, and Trend are used to forecast the unknown variables: AveMaxT, AveMinT, AveTem, nonlintrend18, nonlintrend19, nonlintrend20, and nonlintrend21. The forecasted predictor variables are then used to forecast electricity demand at hours 18:00, 19:00, 20:00, and 21:00, shown in Figure 4.6. Figure 4.6 shows the electricity demand superimposed with forecasts for hours 18:00 and 19:00 using the AQR model and for 20:00 and 21:00 using the LQR model. There is strong seasonality in all of the plots, with upward trends. It is clear from the overall upward drifts that the time series indicate increasing trends.

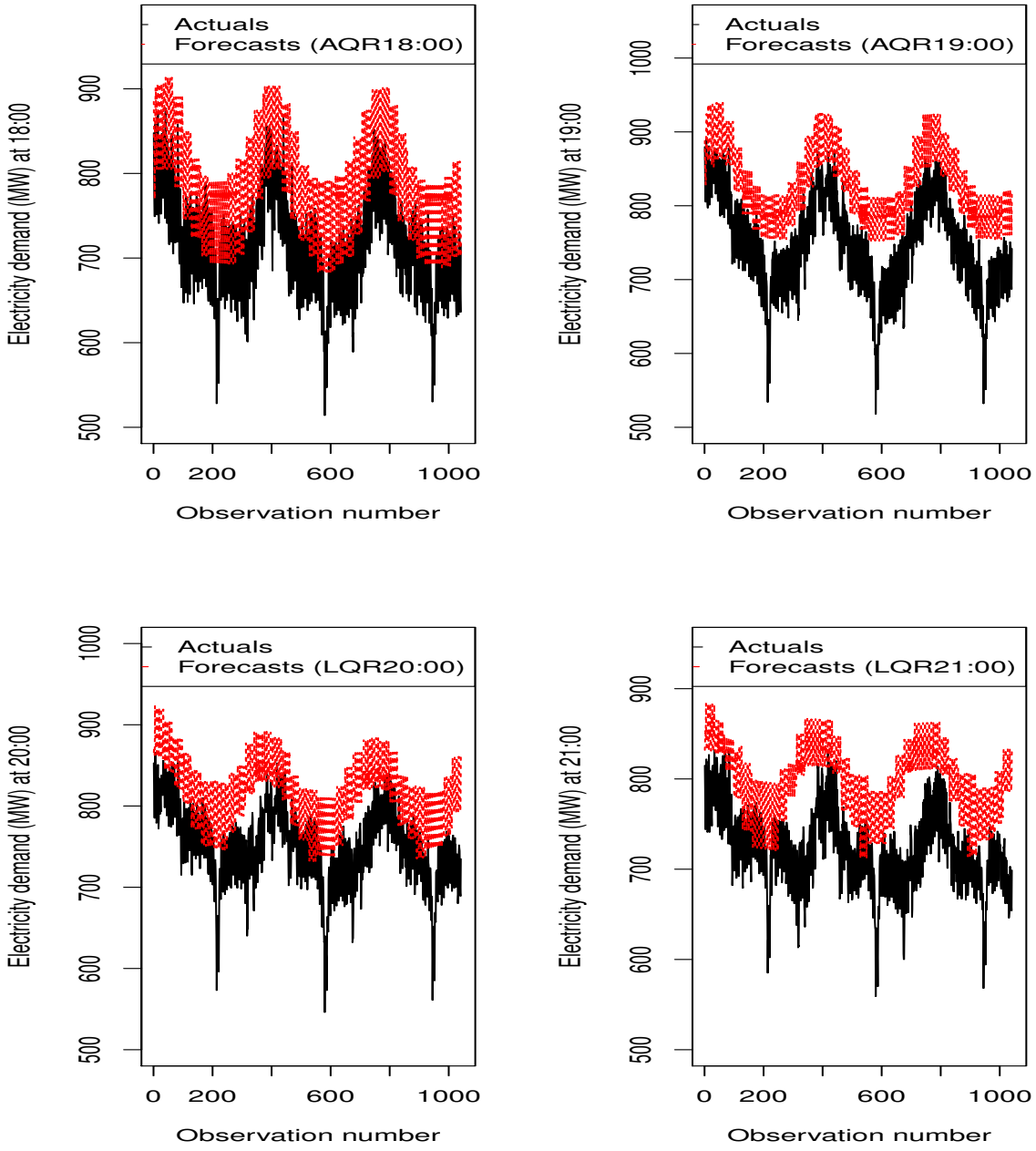

Figure 4.6: Plot of actual demand and forecasts at hours 18:00, 19:00, 20:00, and 21:00 (Operational forecasts).

4.2.3 Unit commitment results

Table 4.3 shows the base load demand stations, i.e. CSP and PV. The average production cost, shown in column 2 of Table 4.3, represented the fuel cost in this study. Columns 3 and 4 represent the megawatts' minimum and maximum production levels, respectively. Data for peaking stations, i.e. hydroelectric and wind power, are given in Table 4.4, which is similar to Table 4.3. The fuel costs data is from (Gxasheka et al., 2018). The out-of-sample forecasts for the first seven days of April 2014 obtained using the models for hours 18:00, 19:00, 20:00 and 21:00 are presented in Table 4.5. The application unit commitment problem is demonstrated using the forecasted demand for Monday, 07 April 2014, shown in bold. UC problem was solved using Lingo version 18. Although the study applied the UC problem for the day 07 April 2014, it is crucial to remember that the process can be used at various times of the day. The bounded variable MILP model is given by:

$$\text{Min TC} = c_1 g_{1d} + \sum_{j=2}^{9} c_j g_{jw} + \sum_{j=10}^{17} c_j g_{jc} + \sum_{j=18}^{36} c_j g_{jp}, \tag{4.2.1}$$

where TC is the total cost to be minimised. The generating units hydroelectric, wind power, CSP and PV are denoted by d, w, c and p respectively. The demand used for hours 18:00, 19:00, 20:00 and 21:00 are $P_D^{18t} = 821.145$, $P_D^{19t} = 894.737$, $P_D^{20t} = 879.355$ and $P_D^{21t} = 816.819$ respectively.

Table 4.3: Base load demand stations.

Unit	Min Ave prod cost c_i (rands)	Min (MW)	Max (MW)
g10c	3753	0	50
g11c	3753	0	100
g12c	3753	0	36.652
g13c	3753	0	100
g14c	3753	0	100
g15c	3753	0	100
g16c	3753	0	100
g17c	3753	0	100
g18p	2518	0	60.085
g19p	2518	0	9.65
g20p	2518	0	9.65
g21p	2518	0	10
g22p	2518	0	19.9
g23p	2518	0	19.93
g24p	2518	0	48.3
g25p	2518	0	9.7
g26p	2518	0	40
g27p	2518	0	75
g28p	2518	0	45.4
g29p	2518	0	64
g30p	2518	0	75
g31p	2518	0	74
g32p	2518	0	36.8
g33p	2518	0	75
g34p	2518	0	8.9
g35p	2518	0	75
g36p	2518	0	14.915

Table 4.4: Peaking stations.

Unit	Min Ave prod cost c_i (rands)	Min (MW)	Max (MW)
g1d	1470	0	77.46
g2w	1293	0	73.8
g3w	1293	0	139
g4w	1293	0	96
g5w	1293	0	140
g6w	1293	0	80
g7w	1293	0	140
g8w	1293	0	70.555
g9w	1293	0	70.555

Table 4.5: Out of sample forecasts.

Date	18:00	19:00	20:00	21:00
Tue 01-04-14	816.721	869.6	884.442	827.603
Wed 02-04-14	827.579	862.975	872.569	821.538
Thu 03-04-14	840.839	856.346	859.523	815.473
Fri 04-04-14	814.269	858.669	835.682	792.56
Sat 05-04-14	778.888	848.844	825.485	786.495
Sun 06-04-14	735.893	835.653	818.171	780.43
Mon 07-04-14	**821.145**	**894.737**	**879.355**	**816.819**

Load balance constraints:

$$P_D^{ht} = c_1 g_{1d} + \sum_{j=2}^{9} c_j g_{jw} + \sum_{j=10}^{17} c_j g_{jc} + \sum_{j=18}^{36} c_j g_{jp}, \tag{4.2.2}$$

The demand used for hours 18:00, 19:00, 20:00 and 21:00 are $P_D^{18t} = 821.145$, $P_D^{19t} = 894.737$, $P_D^{20t} = 879.355$ and $P_D^{21t} = 816.819$ respectively. According to Lyons and Gross (2015), utility's reserve margin is usually assigned by its

regulatory agency between 10% and 15%. In this study we use a 14% reserve margin and get:

$$P_D^{18t} = 0.14 \times 821.145 = 114.960 \text{ MW}$$
$$P_D^{19t} = 0.14 \times 894.737 = 125.263 \text{ MW}$$
$$P_D^{20t} = 0.14 \times 879.355 = 123.110 \text{ MW}$$
$$P_D^{21t} = 0.14 \times 816.819 = 114.355 \text{ MW}$$

The generator power output limits:

$$0 \leq g_{1d} \leq 77.46 \tag{4.2.3}$$
$$0 \leq g_{36p} \leq 14.915 \tag{4.2.4}$$

The optimal solution for each of the four hours on Monday, 07 April 2014, is presented in Table 4.6. The amount of electricity in MW of the generating units is shown in column 2 of Table 4.6, which should be produced to meet the predicted demand while ensuring that the power reserve constraint is met. It also displays the minimal cost of providing electricity every four hours in rands. This study is different from Lebotsa et al. (2018) study because it uses extremely high quantile ($\tau = 0.9999$) forecasts to solve the UC problem. With the economic dispatching and UC scheduling of electricity, power utility companies like Eskom can use this procedure.

Table 4.6: Optimal solution using bounded variable MILP.

	18:00 (R2,067,643)		19:00 (R2,483,463)		20:00 (R1,137,006)		21:00 (R1,069,858)	
Variable	Soln	RC	Soln	RC	Soln	RC	Soln	RC
g1d	-	-	-	-	-	-	77.46	0
g2w	-	-	-	-	73.8	0	73.8	0
g3w	-	-	-	-	139	0	139	0
g4w	-	-	-	-	96	0	96	0
g5w	-	-	-	-	140	0	140	0
g6w	-	-	-	-	80	0	80	0
g7w	-	-	-	-	140	0	140	0
g8w	-	-	-	-	140	0	70.555	0
g9w	-	-	-	-	70.555	0	-	-
g10c	0	1235	50	0	0	2460	-	-
g11c	0	1235	100	0	0	2460	-	-
g12c	0	1235	36.652	0	0	2460	-	-
g13c	0	1235	-	-	0	2460	-	-
g14c	0	1235	-	-	0	2460	-	-
g15c	0	1235	-	-	0	2460	-	-
g16c	0	1235	-	-	-	-	-	-
g17c	0	1235	-	-	-	-	-	-
g18p	75	0	60.085	0	-	-	-	-
g19p	9.65	0	75	0	-	-	-	-
g20p	9.65	0	28	0	-	-	-	-
g21p	10	0	75	0	-	-	-	-
g22p	19.9	0	75	0	-	-	-	-
g23p	19.93	0	75	0	-	-	-	-
g24p	48.3	0	75	0	-	-	-	-
g25p	9.7	0	75	0	-	-	-	-
g26p	75	0	40	0	-	-	-	-
g27p	75	0	75	0	-	-	-	-
g28p	45.4	0	55	0	-	-	-	-
g29p	64	0	-	-	-	-	-	-
g30p	75	0	-	-	-	-	-	-
g31p	74	0	-	-	-	-	-	-
g32p	36.8	0	-	-	-	-	-	-
g33p	75	0	-	-	-	-	-	-
g34p	8.9	0	-	-	-	-	-	-
g35p	75	0	-	-	-	-	-	-
g36p	14.915	0	-	-	-	-	-	-

4.3 Conclusion

Using Northern Cape data, this study applied LQR, NLQR, and AQR models to peak electricity demand forecasting at extreme quantile ($\tau = 0.9999$). The paper's thrust was on peak hourly electricity demand at an extremely high quantile ($\tau = 0.9999$). Lasso was used for variable selection. For hours 18:00 and 19:00, the best-fitting models were found to be AQR models, whereas LQR models were the best-fitting models for hours 20:00 and 21:00. The best-fitting models were used to forecast the demand for each hour. Operational forecasts were also done for each of the four hours. Out of sample forecasts given in Table 4.5 from the four hours were then utilised as inputs in solving the UC problem. The Lingo version 18 was used to solve optimisation models. The out-of-sample forecasts for Monday, 07 April 2014, were 821.145 MW, 894.737 MW, 879.355 MW and 816.819 MW for hours 18:00, 19:00, 20:00 and 21:00, respectively, with the corresponding optimal minimal costs of R2,067,643.00, R2,483,463.00, R1,137,006.00 and R1,069,858.00. The UC problem results showed that using all the hydroelectric generating units, wind power, CSP, and PV is less costly. These were all selected as part of the optimal solution. This study's main contribution has been the development of models for forecasting hourly extreme peak electricity demand. These results could be helpful to system operators in the energy sector who have to maintain the minimum cost by scheduling and dispatching electricity during peak hours when the grid is constrained due to peak load demand.

Chapter 5

Estimation of extreme quantiles of global horizontal irradiance: A comparative analysis using an extremal mixture model and a generalised additive extreme value model

5.1 Introduction

Grid energy management is challenging due to the intermittent nature of solar power. Accurate forecasts are essential in integrating solar power into the electricity grid effectively. To achieve sustainability, it is important to generate power from clean sources. Energy policymakers are establishing more penetration of clean energy onto the grid, such as solar and wind power. Lack of accurate renewable energy forecasts and inconsistency in renewable power generation are the limitations for using renewable energy sources (RES) to produce power for supplying peak energy demand (Osório et al., 2015). Ac-

curate forecasting models are useful to system operators and market players in solving problems created by the integration of RES and are also viable to handle their variability and uncertainty (Gonçalves et al., 2021). However, this study focuses on modelling and forecasting global horizontal irradiance (GHI) at extremely high quantiles.

Solar power is one of the preferable RES to be integrated into the grid because its contract price is lower since it depends on primary sources (Mpfumali et al., 2019). The energy sector needs to take the direction of exploring solar energy as it is one of the cleanest energy sources (Sun et al., 2018). Accurate solar power forecasts are useful for economical operation dispatch, optimal unit commitment, and national grid stability. It also reduces the uncertainties of solar energy sources and assures safety and easier grid management. Lack of accurate forecasts of solar power generation creates problems for grid energy management (Mpfumali et al., 2019; Antonanzas et al., 2016).

Solar irradiance forecasting is required to guide the grid operators to apply the relevant procedure to optimise electricity production and reduce production costs. Solar irradiance forecasts are used in an electrical grid to integrate photovoltaic (PV) power (Antonanzas et al., 2016). Power system operators can have problems because of inaccurate extreme quantiles of GHI forecasts. Accurate extreme forecasts of GHI quantiles help decision-makers make high-risk decisions and regulatory frameworks requiring high-security levels. However, accurate forecasts are important for the effective operation of the electrical grid when solar energy is integrated. It has been shown that

solar power forecasting is essential in planning operations in the energy sector, such as unit dispatch and renders schedules of power production from renewable energy sources for the next hours or days. These forecasts are also used to determine the amount of solar power that will be integrated into the grid in the following hours or days. The purpose of producing new forecasting models for solar power is to provide accurate forecasts (Zieher et al., 2015).

5.2 Empirical results

The GHI data used in this study covered January 1 to December 31, 2020, yielding a sample size of 4083 observations. The data is obtained from the SAURAN website (http://sauran.net). In this study, the station of focus is the University of Venda radiometric station found in an inland location at Vuwani Science Research Centre in South Africa, Limpopo province. It is located at coordinates -23.13100052, 30.42399979, and its elevation is 628 m. Figure 5.1 shows a picture of the pyranometer at the University of Venda radiometric station. The study seeks to predict extremely high quantiles of GHI using R statistical software.

Figure 5.1: UNIV – USAid Venda station. Source: https://sauran.ac.za/.

5.2.1 Exploratory data analysis

The summary statistics of hourly GHI are given in Table 5.1. Since the mean and median of the GHI data are not equal, the distribution is not normally distributed. The non-normality of the distribution of GHI data is confirmed by the skewness and kurtosis values which are 0.5734277 and -0.9057017, respectively.

Table 5.1: Summary statistics for GHI.

Mean	Median	Min	Max	Std	Skewness	Kurtosis
355.2654	278.7552	0.0005	1121.8900	316.1626	0.5734	-0.9057

Figure 5.2 presents the plot of GHI together with histogram and box plots.

Figure 5.3 displays the normal Q–Q plot of GHI. The plots show seasonality with upward trends. It indicate that the data is not normally distributed. These plots suggest that the GHI data does not follow a Gaussian distribution. A plot of hourly GHI superimposed with a non-linear trend is shown in Figure 5.4, and this best fit was determined based on "mgcv" developed by Wood (2017).

5.2.2 Forecasting results

An AQR model combines GAM and QR similar to the one discussed in Section 3.2.3. The AQR model is defined by (Gaillard et al., 2016; Fasiolo et al., 2017, 2020a,b).

$$y_{t,\tau} = \sum_{i=1}^{p} s_{i,\tau}(x_{ti}) + \varepsilon_{t,\tau}; \quad \tau \in (0,1), \tag{5.2.1}$$

where $y_{t,\tau}$ presents a GHI at time $t = 1,..,n$ at quantile τ, $s_{i,\tau}$ denote the smooth functions and $\xi_{t,\tau}$ denotes the error term. In this study, we are going to consider two scenarios: one AQR model (5.2.2) (AQR-1), which has one linear covariate t and two AQR model (5.2.3) (AQR-2) which has one linear covariate t and temperature variable. The two models are written as follows:

$$q_{Y|X}(\tau) = \sum_{t=1}^{n} \rho_\tau\big(y_{t,\tau} - s(t)\big), \quad \tau \in [0.95, 1), \tag{5.2.2}$$

$$q_{Y|X}(\tau) = \sum_{t=1}^{n} \rho_\tau\big(y_{t,\tau} - s(t) - s(\text{temp})\big), \quad \tau \in [0.95, 1), \tag{5.2.3}$$

The models considered in this study are the SPEM, GAEV, BM, AQR-1 and AQR-2. The R-packages used in this study are "extremefit" developed by

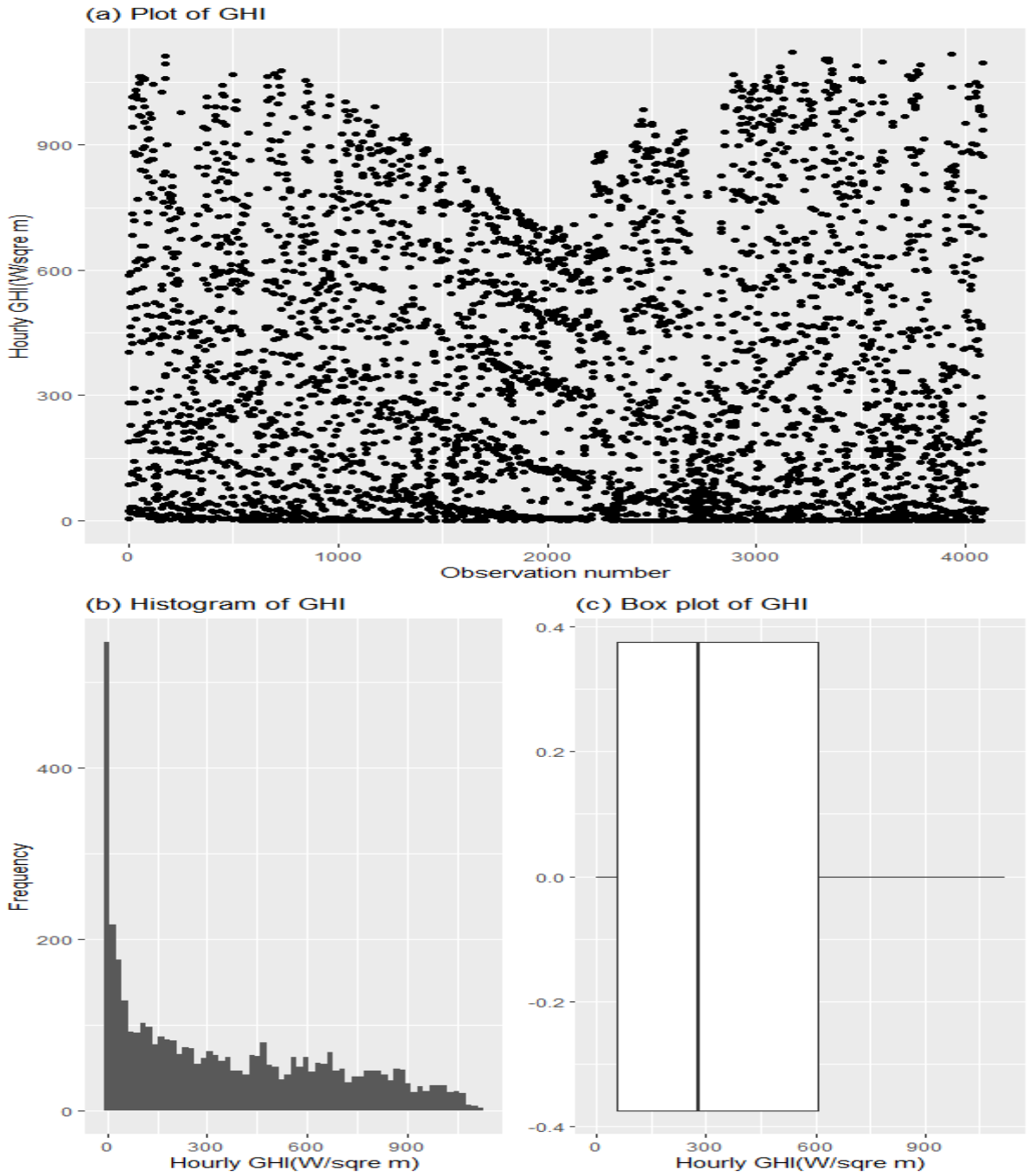

Figure 5.2: Diagnostic plots for hourly GHI from 01 January 2020 to 31 December 2020.

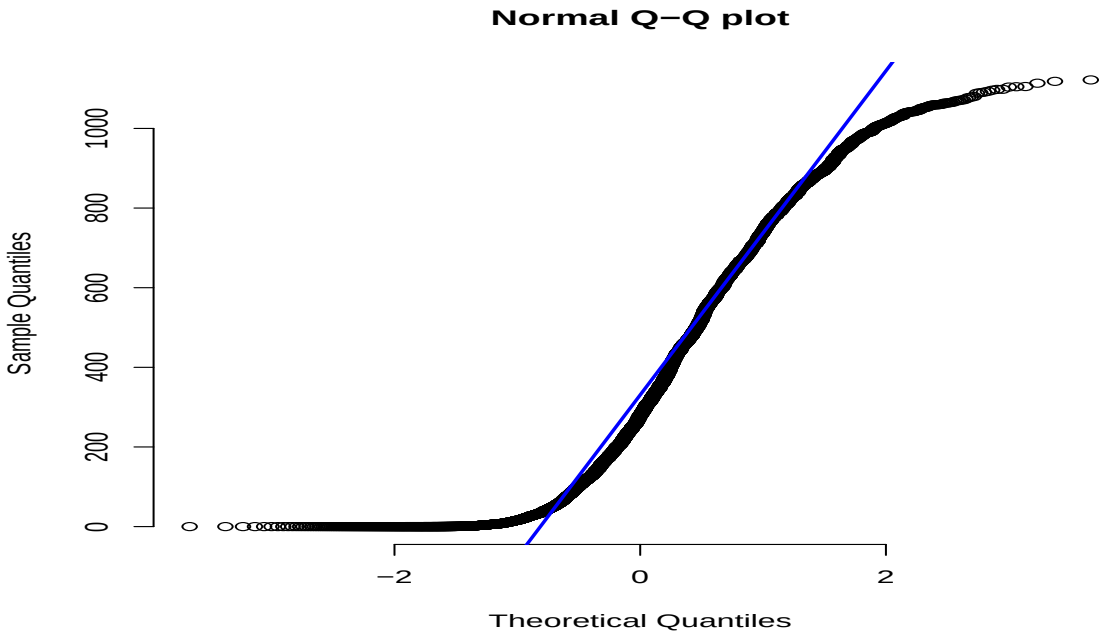

Figure 5.3: Normal Q–Q plot for hourly GHI.

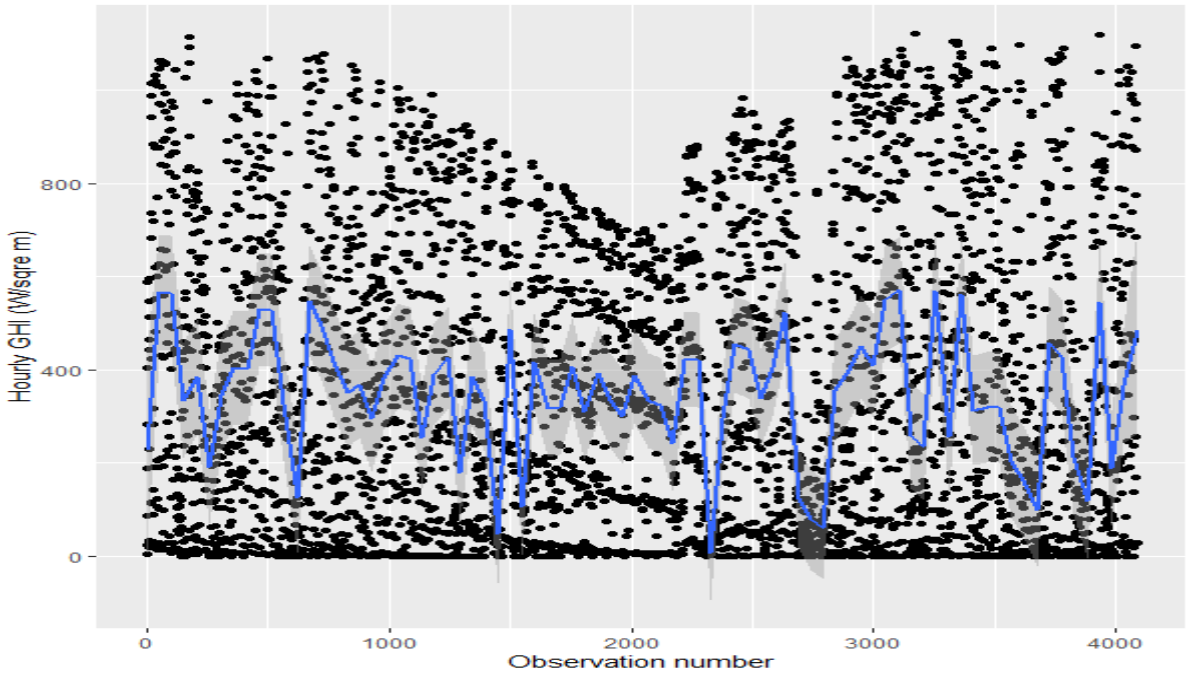

Figure 5.4: Plots of hourly GHI from 01 January 2020 to 31 December 2020 superimposed with a non-linear trend.

Durrieu et al. (2016), which is used for the SPEM, "evgam" developed by Youngman (2020), which is used for the GAEV model and "qgam" developed by Fasiolo et al. (2020b), which is used for AQR models. The quantiles considered in this study are $\tau = 0.95$, 0.97, 0.99, 0.999 and 0.9999. All the data is used for predictions. The plots of actual hourly GHI and forecasts of GHI at quantiles ($\tau = 0.95$, 0.97, 0.99, 0.999 and 0.9999) using SPEM, GAEV, BM, AQR-1 and AQR-2 models are given in Figures 5.5 – 5.7 respectively. It is shown on the right side of Figure 5.5 and the left side of Figure 5.6 that each model's quantile predictions follow the hourly GHI data remarkably well at a high quantile. Using the SPEM model on the left side of Figure 5.5, the extreme value forecasts follow the actual demand extreme well at quantiles ($\tau = 0.95$, 0.97, 0.99 and 999) except at quantile $\tau = 0.9999$, which contains extreme outliers. Consequently, the SPEM model performs poorly at extremely high quantiles. The right side of Figure 5.6 shows the results from the AQR-2 model, which shows that some actuals are greater than forecasts at quantiles ($\tau = 0.95$, 0.97 and 0.99). The AQR-2 model does not produce useful forecasts at extremely high quantiles ($\tau = 0.999$ and 0.9999). The simple average (AVG) and median (MED) methods combine forecasts for the SPEM, GAEV, BM, AQR-1 and AQR-2 models at each quantile. The plots of actual hourly GHI and forecasts of GHI using AVG and MED methods at quantiles ($\tau = 0.95$, 0.97, 0.99, 0.999 and 0.9999) are given in Figure 5.8. As shown by the left side of Figure 5.8, the extreme value corresponds to the depth of the actual demand, except for the extreme outliers at quantile ($\tau = 0.9999$).

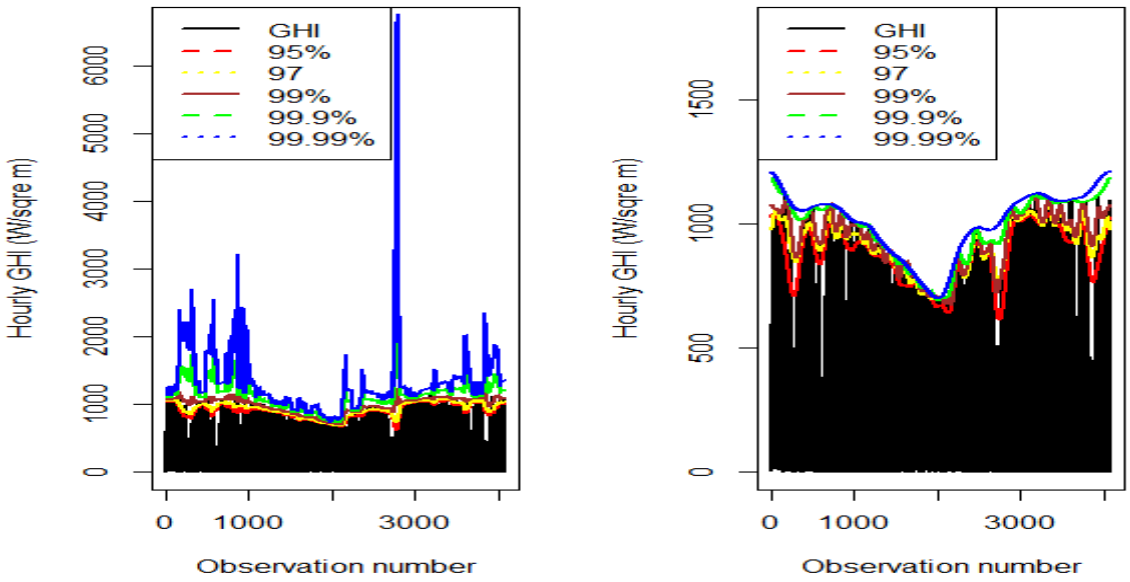

Figure 5.5: Plots of actual hourly GHI and forecasts of GHI at high quantiles using SPEM (left-side) and GAEV (right-side) models.

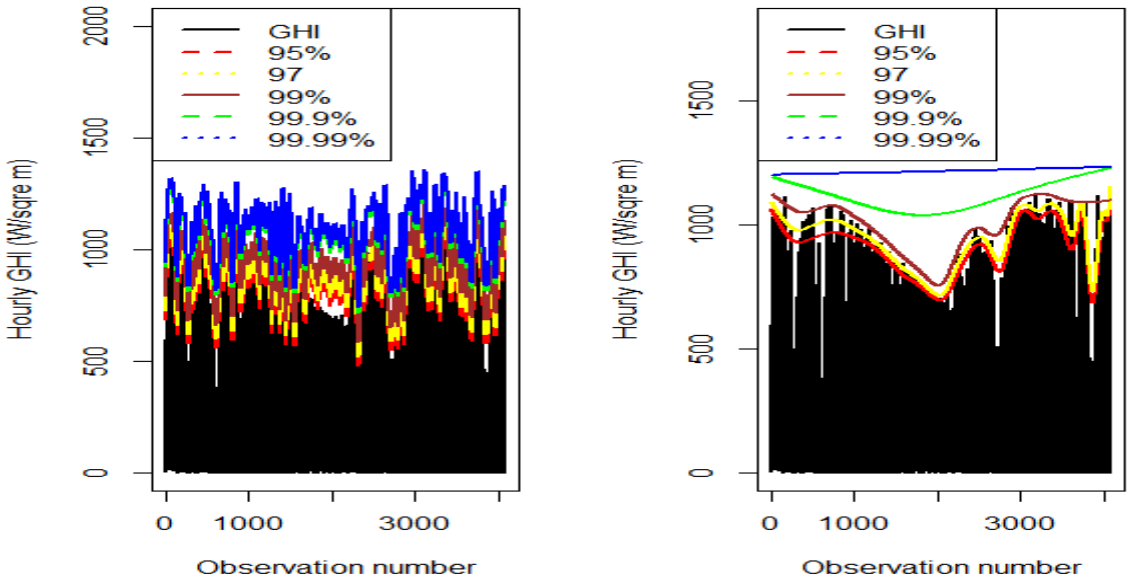

Figure 5.6: Plots of actual hourly GHI and forecasts of GHI at high quantiles using BM (left-side) and AQR-1 (right-side) models.

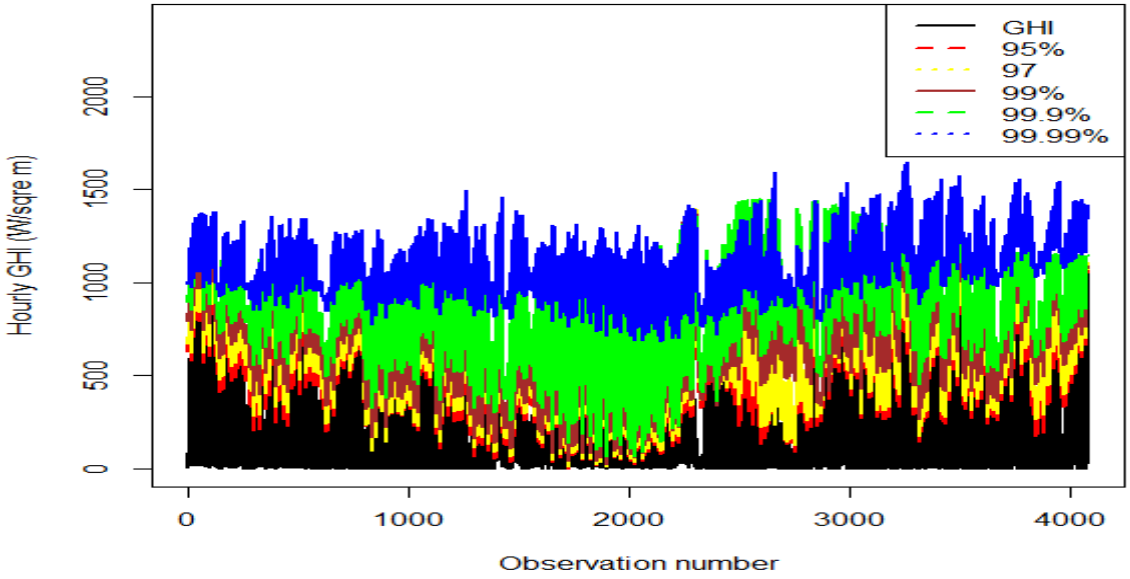

Figure 5.7: Plots of actual hourly GHI and forecasts of GHI at high quantiles using AQR-2 model.

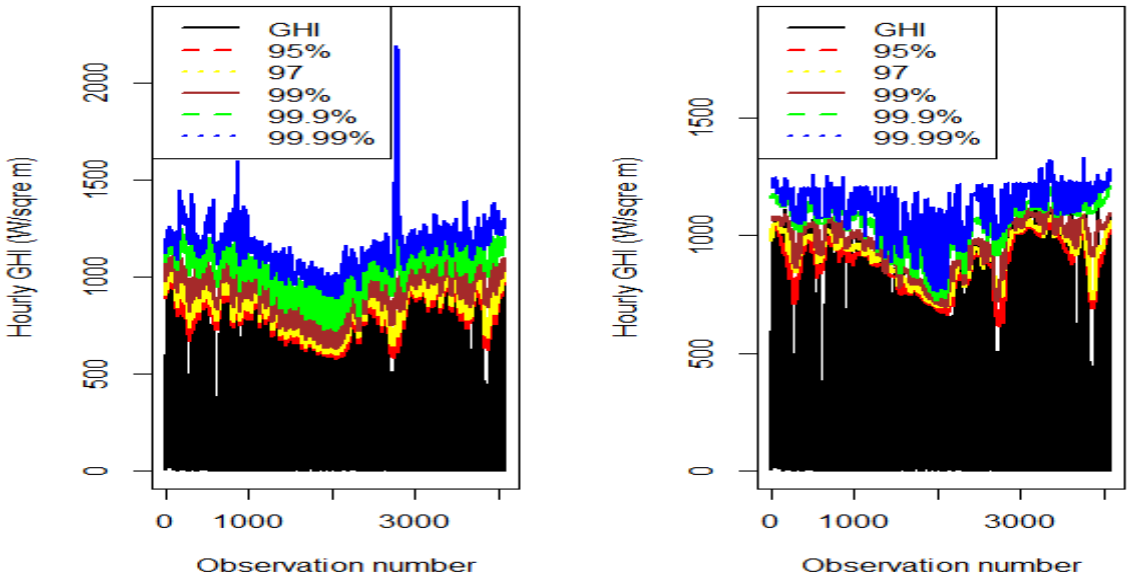

Figure 5.8: Plots of actual hourly GHI and forecasts of GHI at high quantiles using AVG (left-side) and MED (right-side) model.

5.2.3 Comparative Analysis

Table 5.2 gives a summary of the values of the evaluation metrics (PLF and CRPS) for the proposed models. At quantile levels, $\tau = 0.95$, 0.97, 0.99 and 0.999, AQR-2 is the most accurate model based on the PLF, and $\tau = 0.9999$, GAEV is the most accurate model based on the PLF and CRPS.

Table 5.2: Model comparisons using PLF and CRPS.

Accuracy	Model	$\tau = 0.95$	$\tau = 0.97$	$\tau = 0.99$	$\tau = 0.999$	$\tau = 0.9999$
PLF	SPEM	29.4337	18.2303	6.3971	0.7809	0.0976
	GAEV	28.7070	17.6628	5.9598	0.6294	**0.0654**
	BM	27.5458	17.5389	6.4691	0.7302	0.0744
	AQR-1	29.5307	18.3051	6.5218	0.7615	0.0861
	AQR-2	**20.7937**	**13.6359**	**5.1719**	**0.5856**	0.0763
	AVG	26.3300	16.5366	5.9588	0.6952	0.0798
	MED	28.3282	17.5617	6.1030	0.7024	0.0769
CPRS	SPEM	27226.53	27255.53	27315.11	27464.02	27658.15
	GAEV	27203.16	27240.34	27264.59	27311.81	**27333.55**
	BM	27167.32	27207.16	27271.90	27403.37	27426.28
	AQR-1	27235.74	27271.83	27333.23	27444.72	27543.50
	AQR-2	**27045.44**	**27089.42**	**27177.64**	**27269.47**	27445.83
	AVG	27175.64	27212.85	27272.49	27378.68	27481.46
	MED	27210.31	27245.64	27286.57	27385.88	27452.12

Figure 5.9 shows the plots of actual hourly GHI and forecasts of GHI from the best-fitting model AQR-2 at quantiles $\tau = 0.95$, 0.97, 0.99 and 0.999. Figure 5.10 also shows the plots of actual hourly GHI and forecasts of GHI from the best-fitting model GAEV at quantile $\tau = 0.9999$. The Figures demonstrate some under-prediction of forecasts from best-fitting models.

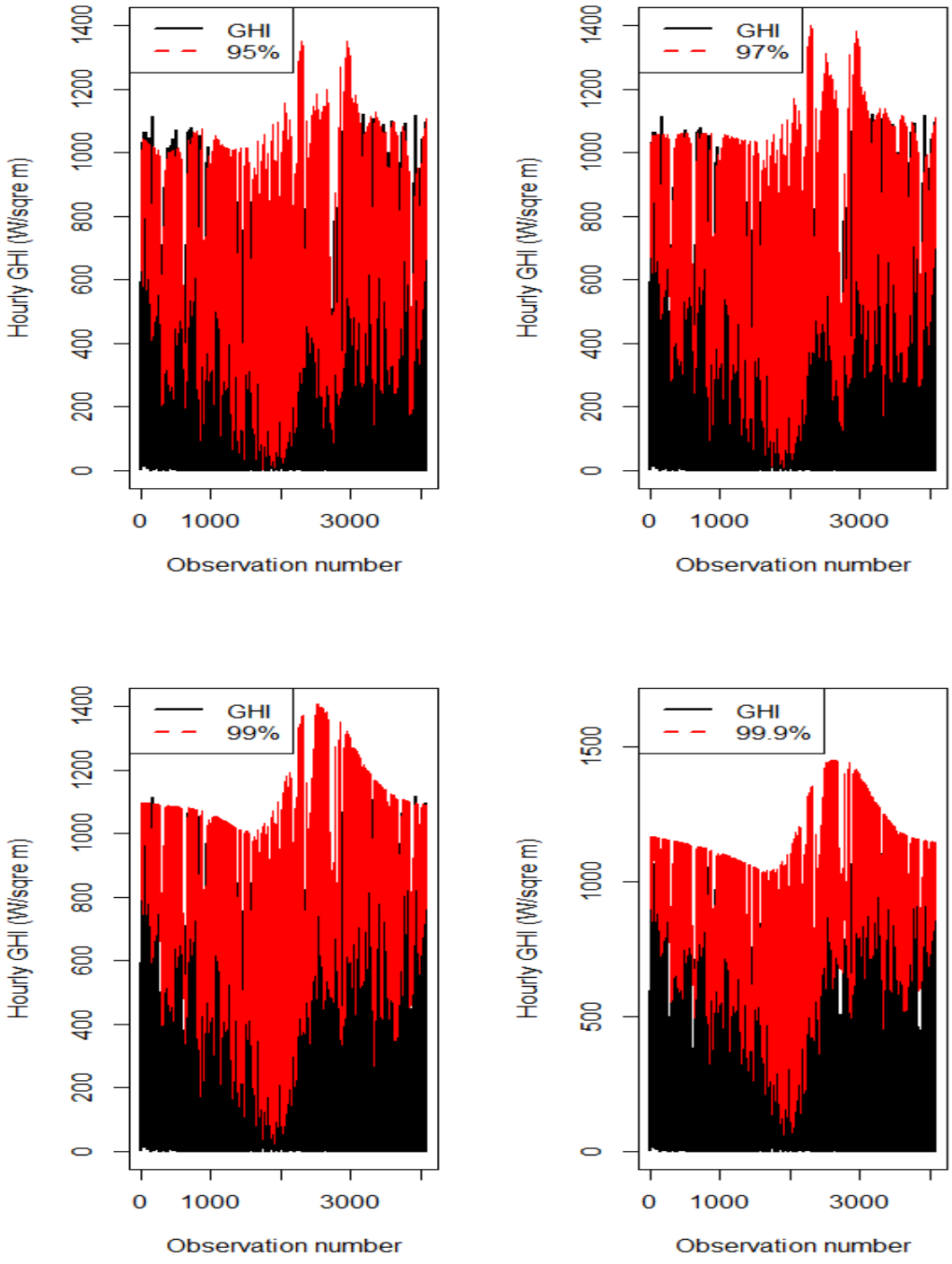

Figure 5.9: Plots of actual hourly GHI and forecasts of GHI using AQR-2 model at $\tau = 0.95, 0.97, 0.99$ and 0.999.

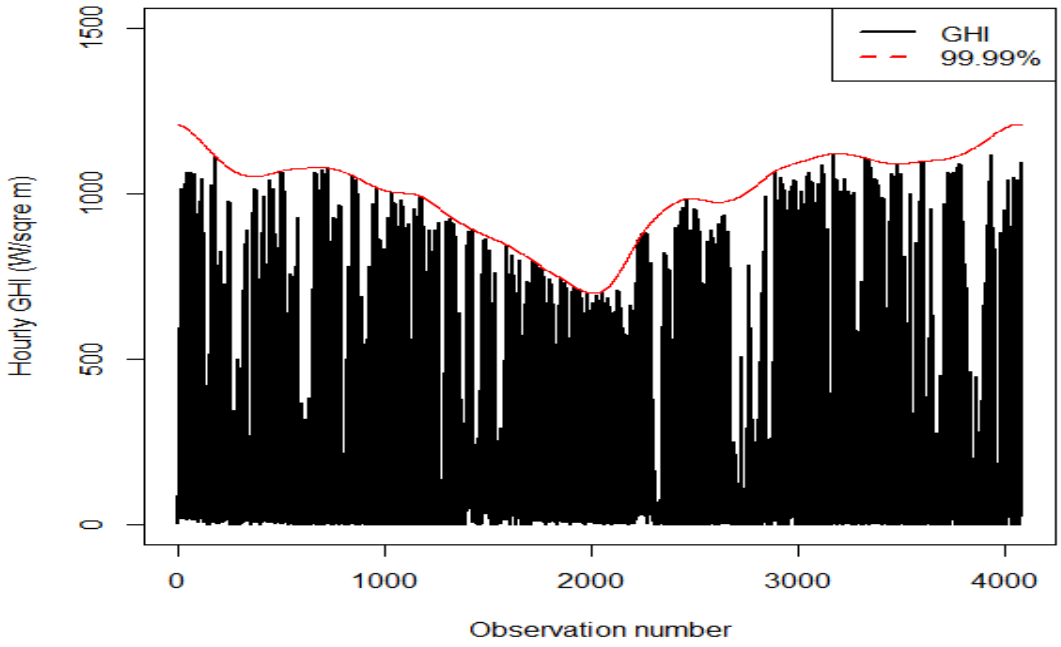

Figure 5.10: Plots of actual hourly GHI and forecasts of GHI at high quantiles using GAEV model at $\tau = 0.9999$

5.2.4 Discussion of results

This study was motivated by previous studies such as Mpfumali et al. (2019); Mutavhatsindi et al. (2020), among others. The focus of the study was on a comparative analysis of SPEM, GAEV, BM, AQR-1 and AQR-2 models in predicting extremely high hourly GHI using South African data. The forecasts from SPEM, GAEV, BM, AQR-1 and AQR-2 models were combined using AVG and MED methods to develop AVG and MED forecasts. The PLF and CRPS were used to measure the accuracy of models at each quantile, including AVG and MED forecasts.

Based on the PLF and CRPS, the best-fitting model for quantiles $\tau = 0.95$, 0.97, 0.99 and 0.999 was found to be AQR-2, while for the 0.9999-quantile, the best model is GAEV. The most accurate model was selected based on the smallest value of PLF and CRPS. It was revealed that fitting a model (i.e. AQR-2) including temperature variable improves the forecasts because the temperature is the main predictor of GHI. System operators could benefit from knowing the greatest solar power that can be integrated into the national grid by predicting extremely high quantiles of GHI solar power. Accurate forecasts of GHI produced by this study are useful for system operators and decision-makers in effectively balancing the demand and supply of environmentally friendly electricity that secures the country's future economic prosperity.

5.3 Conclusion

This study compared SPEM, GAEV, BM, AQR-1, AQR-2, AVG and MED models to predict extremely high hourly GHI using data from the University of Venda radiometric station. The thrust of the study was on hourly GHI at extremely high quantiles, i.e. $\tau = 0.95$, 0.97, 0.99, 0.999 and 0.9999. The AQR-2 was the most accurate model for quantiles ($\tau = 0.95$, 0.97, 0.99 and 0.999), and the GAEV model was the most accurate model for quantile ($\tau = 0.9999$). The best-fitting models were also used to forecast the GHI for each quantile. The findings of this study could be helpful to decision-makers from power utilities in the optimal integration of the electricity grid of the energy generated from solar power plants. The predictions from this study are

valuable to power utility decision-makers and system operators when making high-risk decisions and regulatory frameworks that require high-security levels. According to our knowledge, this is the first application to conduct a comparative analysis of the proposed models using South African solar irradiance data.

Chapter 6

Extremal Dependence Modelling of Global Horizontal Irradiance with Temperature and Humidity: An Application Using South African Data

6.1 Introduction

The interaction between global horizontal irradiance (GHI) and temperature helps determine the maximum amount of solar power generated. As temperature increases, GHI increases up until an increase at a decreasing rate occurs, and then it decreases. Therefore, system operators need to know the maximum possible solar power that can be generated. A sustainable energy supply can be achieved through renewable energy, such as solar energy, which is still underused. Solar energy is environmentally friendly since there are no greenhouse emissions when electricity is generated using this resource. This reduces the effects of climate change. The challenge in using renewable

energies such as solar is that they are highly intermittent (Alresheedi and Al-Hagery, 2020).

Solar energy is projected to make up a significant part of the world's energy supply in the future. Accurate forecasting is needed to model intermittent solar energy spatially and temporally. Integrating renewable energy sources into the grid is likely to be one of the most significant challenges of the future world's energy supply (Heinemann et al., 2006). Solar power forecasting is essential for operating power systems and adds ancillary services during times of high penetration of renewable energy to avoid variability and uncertainty problems (Alresheedi and Al-Hagery, 2020). Due to its intermittence, instability and randomness, it is not easy to produce accurate, reliable forecasts of solar irradiance that can be used as a guide for operating modern smart grids for electricity generation and distribution. The amount of power produced by photovoltaic (PV) systems is the same as the global horizontal irradiance (GHI) that strikes the earth's surface. GHI depends on the weather to produce considerable power, so understanding the weather is essential (Gao et al., 2020).

Solar power generation is primarily affected by meteorological factors such as solar radiation, temperature, wind pressure and humidity. Since GHI depends on meteorological variables, it is important to determine the extremal dependence and correlation between GHI with temperature and relative humidity. It is also important to determine the temperature (value) which gives us the maximum GHI. This is important as it enables us to maintain

the optimal electrical power grid operation, reliability, stability and power dispatching abilities (Mishra et al., 2020).

Some studies have been done on forecasting GHI, although most concentrate on forecasting the means or averages of the distribution. However, this study aims to model the dependence of the extreme GHI on temperature and RH. Understanding solar energy's peak power production requires analysis of GHI at the extreme tails. The weather variables, temperature and RH, which are the main drivers of GHI, are used to improve accuracy.

6.2 Empirical Results

The hourly GHI data used in this study are from the SAURAN website (https://sauran.ac.za/, accessed on 27 November 2021) from 16 November 2015 to 16 November 2021, giving a sample size of n = 27,274 observations. The University of Venda's radiometric station was chosen for this study. It is located at the Vuwani Science Research Centre in South Africa, Limpopo province, at latitude −23.13100052, longitude 30.42399979 and an elevation of 628 m. The R software was used for all the statistical analyses undertaken in this study.

6.2.1 Exploratory Data Analysis

A summary of hourly GHI statistics is presented in Table 6.1. Given that the mean and median are not equal, the distribution of the GHI data does not reflect a normal distribution. Skewness and kurtosis values of 0.5109 and −0.9320, respectively, confirm the non-normality of the distribution of

GHI data. The skewness and kurtosis values in Table 6.1 indicate that the temperature and relative humidity (RH) distributions are non-normal. The kurtosis value reveals that the temperature data distribution is platykurtic and is skewed to the right. However, the distribution of RH data is skewed to the left, and kurtosis is platykurtic. Figure 6.1 displays hourly GHI superimposed. The non-linear trend was estimated based on the generalised additive model discussed in Wood (2017).

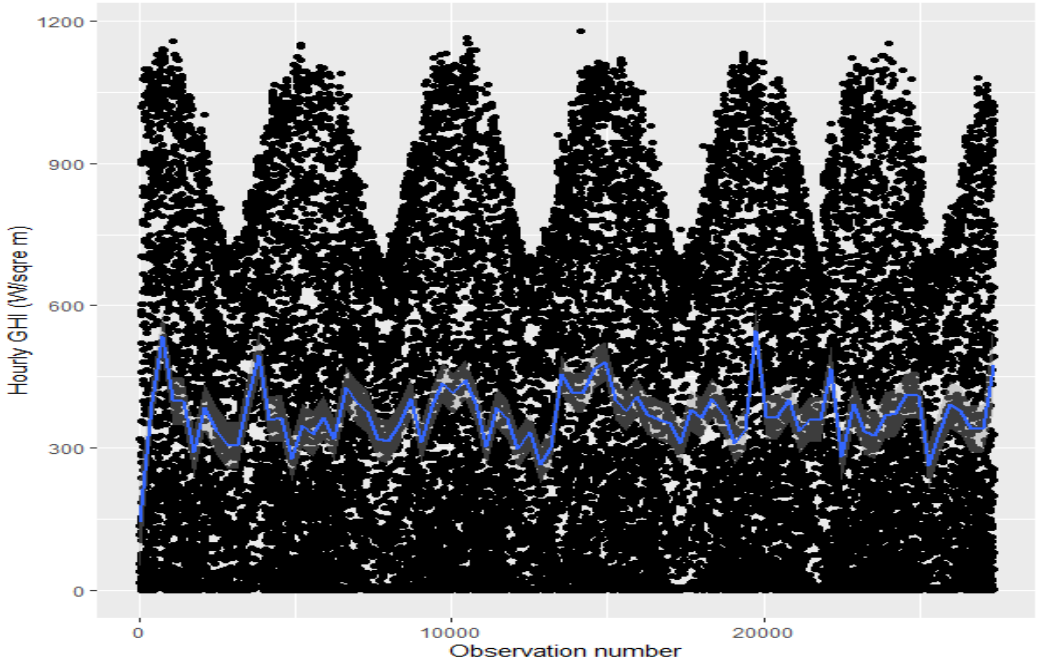

Figure 6.1: Plot of hourly GHI from 16 November 2015 to 16 November 2021 superimposed with a non-linear trend.

Table 6.1: Summary statistics for GHI.

Variable	Mean	Median	Mini	Maxi	Std	Skewness	Kurtosis
GHI	375.642	312.041	0.100	1179.16	314.736	0.511	−0.932
Temp	24.5 °C	24.2 °C	4.9 °C	43.2 °C	5.9 °C	0.1 °C	−0.4 °C
RH	58.365%	57.810%	5.428%	100%	22.607%	−0.015%	−0.962%

6.3 Results for Estimating Extreme Quantiles

6.3.1 MARS Models

This study used the MARS system to model the impact of temperature and RH on GHI. The models are given as follows:

$$\text{GHI} = \beta_0 + \beta_1 \max(0, t_s - \text{T}_t) + \beta_2 \max(0, \text{T}_t - t_w) + \varepsilon_t, \tag{6.3.1}$$

$$\text{GHI} = \beta_0 + \beta_1 \max(0, rh_s - \text{RH}_t) + \beta_2 \max(0, \text{RH}_t - rh_w) + \varepsilon_t, \tag{6.3.2}$$

where $\beta_i, i = 1, 2$ present parameters to be estimated, t_s and rh_s denote reference temperatures and RHs which will be determined using the MARS algorithm, t_w and rh_w represent the peak temperature and RH respectively and error term ε_t (Sigauke and Chikobvu, 2010). In the modelling, marginal GHI increases are estimated for each unit increase in temperature and RH above t_s and rh_s.

Temperature against GHI

The R statistical package "earth" developed by Milborrow (2018) was used to estimate the reference temperatures, t_s and t_w in Equation (6.3.1). Knots were estimated at $t_s = 19$ °C and $t_w = 30$ °C. System operators need to analyse the effects of high temperatures above 30 °C to design, operate and

maintain their power plants. By estimating the parameters and knots of the MARS model for GHI against Temperature, we obtained the following:

$$\hat{\text{GHI}} = 420 + 24 * \max(0, \text{T}_t - 19) - 17 * \max(0, 30 - \text{T}_t) - 18 * \max(0, \text{T}_t - 30) \tag{6.3.3}$$

The Figure 6.2 shows how GHI and temperature superimposed with the estimated MARS model in Equation (6.3.3) are related. Equation (6.3.4) presents a summary of the piecewise linear functions, which the MARS model estimated.

$$\hat{\text{GHI}} = \begin{cases} \hat{\text{GHI}} = 420 - 17 * \max(0, 30 - \text{T}_t), & \text{if } T_t < 19°\text{C} \\ \hat{\text{GHI}} = 420 + 24 * \max(0, \text{T}_t - 19) - 17 * \max(0, 30 - \text{T}_t), & \text{if } 19°C \leq T_{\leq} 30°\text{C} \\ \hat{\text{GHI}} = 420 + 24 * \max(0, \text{T}_t - 19) - 18 * \max(0, \text{T}_t - 30), & \text{if } T_t > 30°\text{C} \end{cases} \tag{6.3.4}$$

We then estimated the gradients of the piecewise linear functions given in Equation (6.3.4). The gradient of the linear model for which $T_t < 19$ °C was found to 17, for 19 °C $\leq T \leq$ 30 °C the gradient was estimated as 41 and for $T_t > 30$ °C it was 24. This means that for 19 °C $\leq T \leq$ 30 °C for a one-degree increase in temperature, there will be an increase of 41 MW.

RH against GHI

The knots of RH were estimated using the MARS model for GHI against RH. They were estimated as $\hat{rh}_s = 48\%$ and $\hat{rh}_w = 84\%$. By estimating the parameters and knots of the MARS model for GHI against RH, we obtained the following:

$$\hat{\text{GHI}} = 540 + 1.1 * \max(0, 48 - \text{RH}_t) - 12 * \max(0, \text{RH}_t - 48) + 6.8 * \max(0, \text{RH}_t - 84) \tag{6.3.5}$$

Equation (6.3.6) presents a summary of the piecewise linear functions estimated by the MARS model for the GHI and RH data.

$$\hat{\text{GHI}} = \begin{cases} \hat{\text{GHI}} = 540 + 1.1 * \max(0, 48 - \text{RH}_t), & \text{if } \text{RH}_t < 48 \\ \hat{\text{GHI}} = 540 - 12 * \max(0, \text{RH}_t - 48), & \text{if } 48 \leq \text{RH}_t \leq 84 \\ \hat{\text{GHI}} = 540 - 12 * \max(0, \text{RH}_t - 48) + 6.8 * \max(0, \text{RH}_t - 84), & \text{if } \text{RH}_t > 84 \end{cases} \tag{6.3.6}$$

Figure 6.3 shows how GHI and RH are superimposed with the estimated MARS model in Equation (6.3.5). The estimated gradients for the piecewise linear functions are as follows: For $\text{RH}_t < 48$, we obtain -1.1; for $48 \leq \text{RH}_t \leq 84$ we obtain -12 and for $\text{RH}_t > 84$, the estimated gradient was found to be -5.

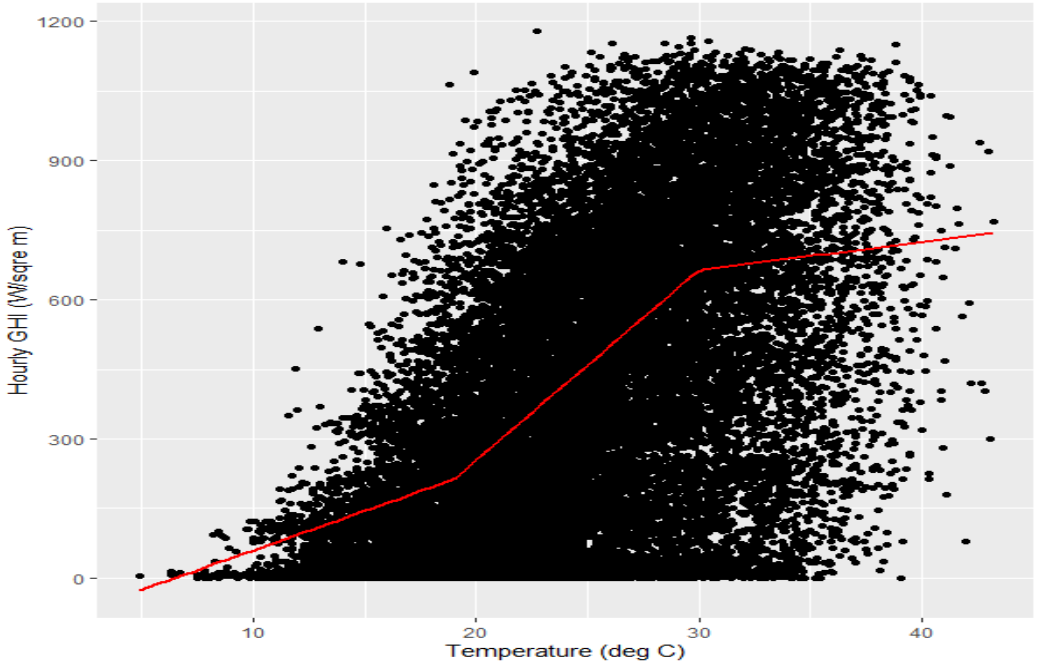

Figure 6.2: Plot of GHI versus Temperature.

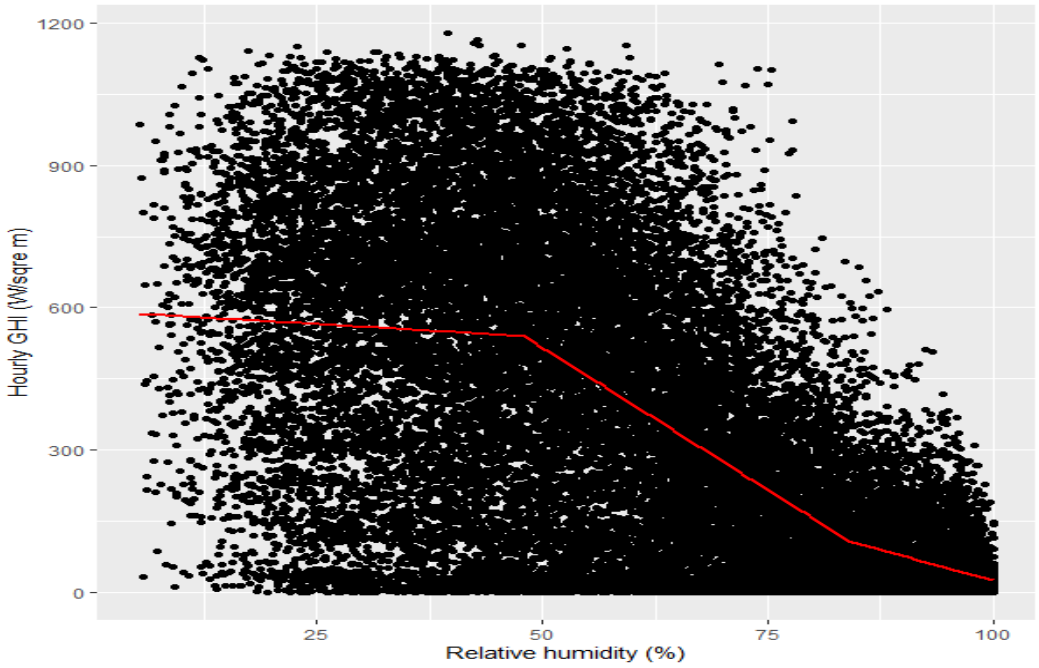

Figure 6.3: Plot of Relative humidity versus GHI.

6.3.2 Threshold Estimation

The R package 'evmix' developed by Hu and Scarrott (2018) was used to estimate thresholds as shown in Figure 6.4 top and bottom panels, respectively. This is based on the extremal models discussed in Section 3.5.1. Figure 6.4 top panel shows a temperature density function of the parametric extreme value mixture model with Weibull distribution for bulk and the tails of the GPD, with bulk and parameterised tail fraction approach, where the vertical lines indicate the thresholds. The thresholds at which the tail fraction would be considered sufficiently high are 32.4 °C and 30.96 °C for bulk and parameterised tail fraction, respectively. Figure 6.4 bottom panel shows an RH density function of the parametric extreme value mixture model with

Weibull distribution for bulk and the tails of the GPD, with bulk and parameterised tail fraction approach. The thresholds at which the tail fraction would be considered sufficiently high are 90.2% for bulk tail fraction and 89.9% parameterised tail fraction. In Figures 6.6 and 6.7, the marginal diagnostic plots for temperature reveal that the points are linear, indicating that the parameterised tail fraction fits the data well. Figures 6.8 and 6.9 given illustrate marginal diagnostic plots of the RH data, which indicate a good fit of the parameterised tail fraction.

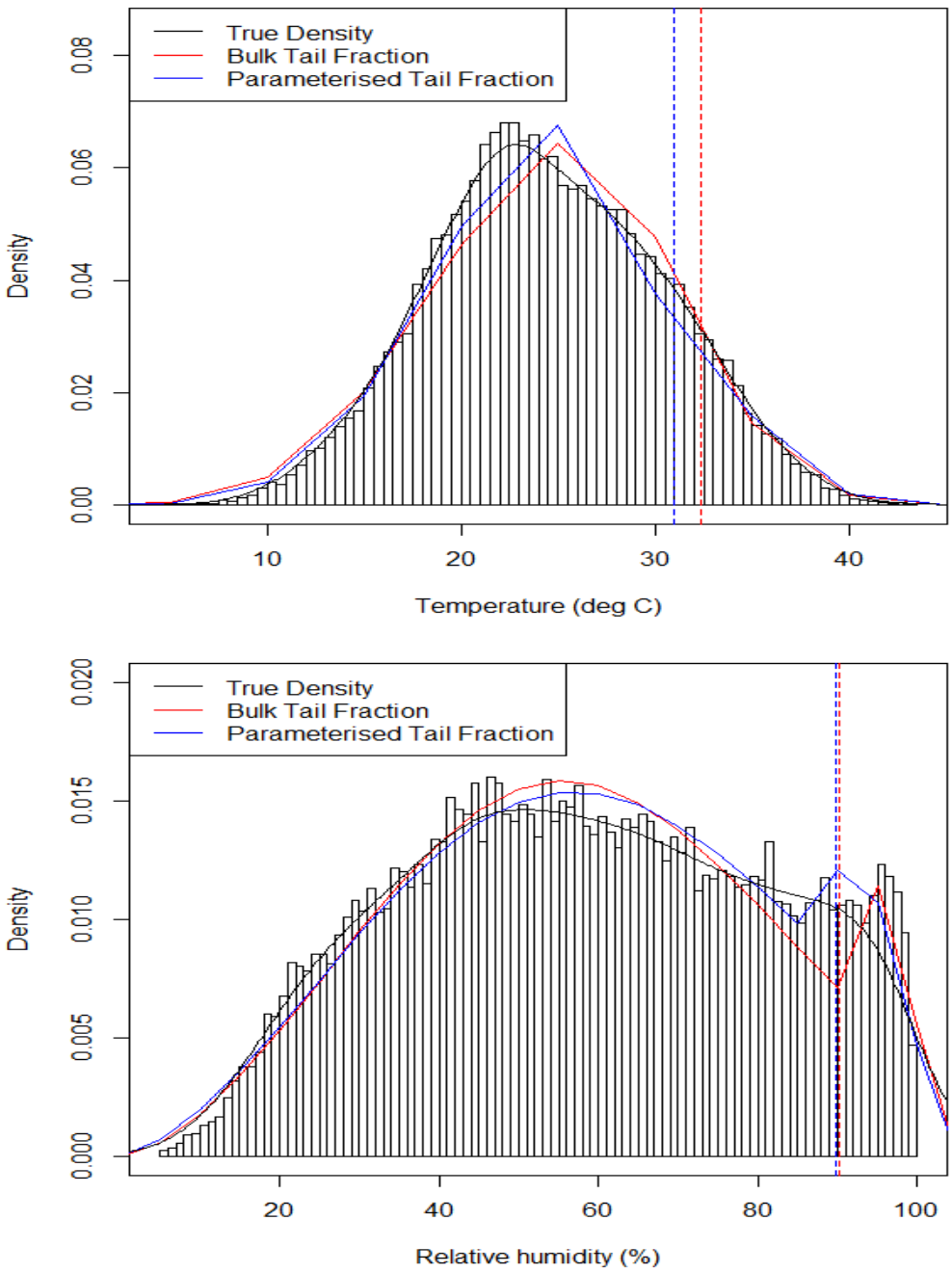

Figure 6.4: **Top panel:** Density function of temperature. **Bottom panel:** Density function of relative humidity.

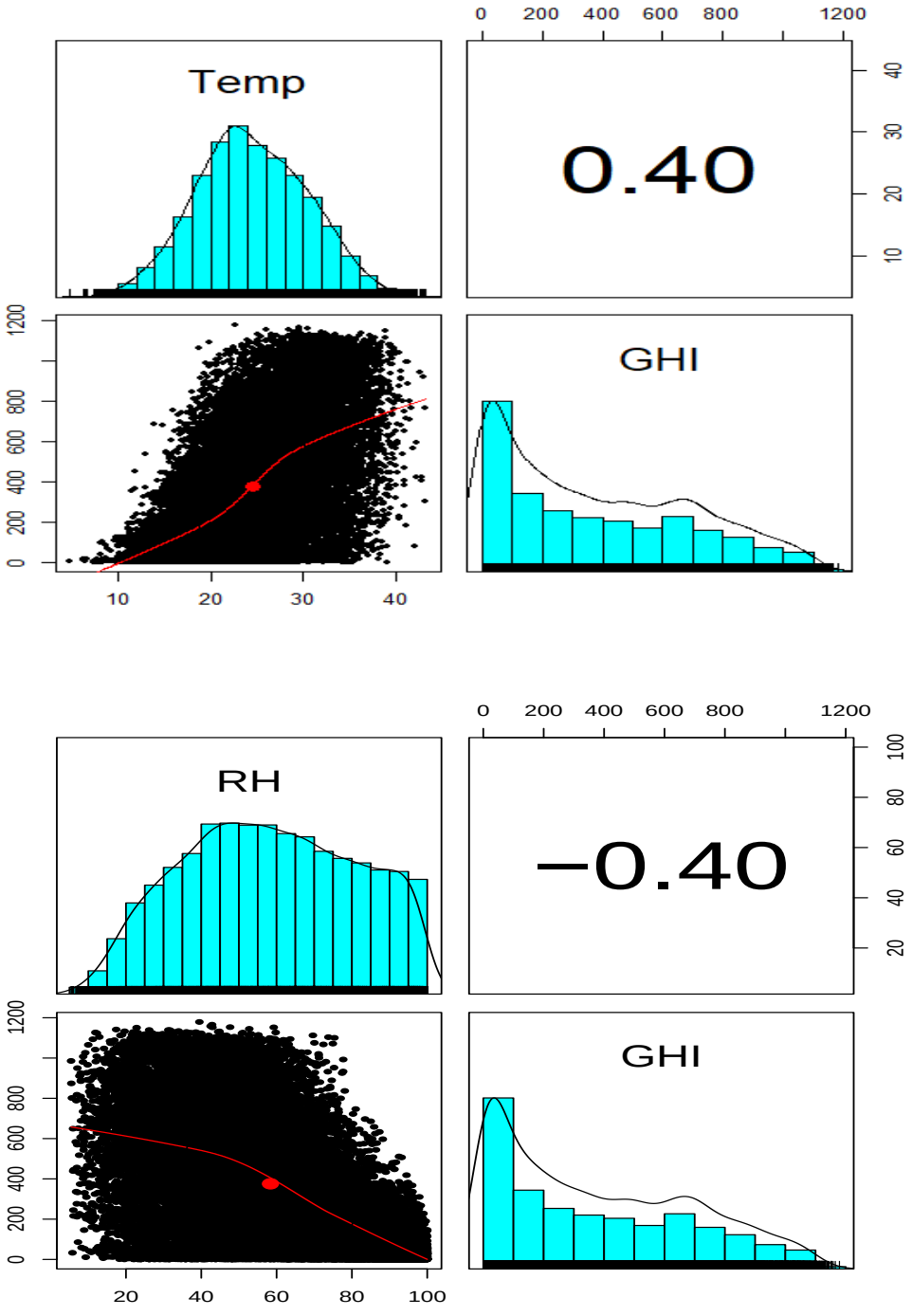

Figure 6.5: **Top panel:** Pairs panel GHI against temperature using the Kendall method. **bottom panel:** Pairs panel GHI and relative humidity.

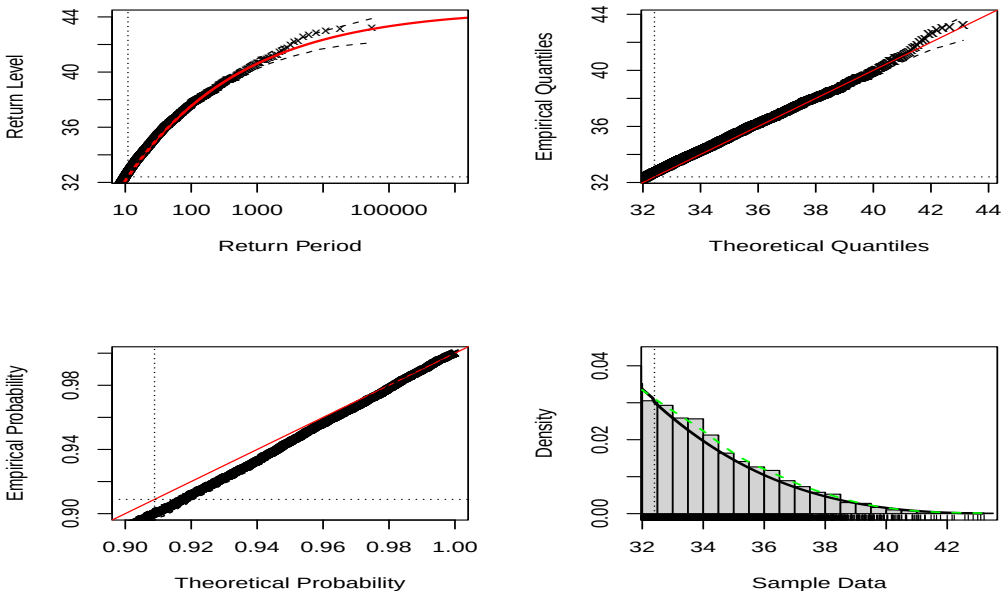

Figure 6.6: Diagnostic plots for fitted Bulk model of Temperature.

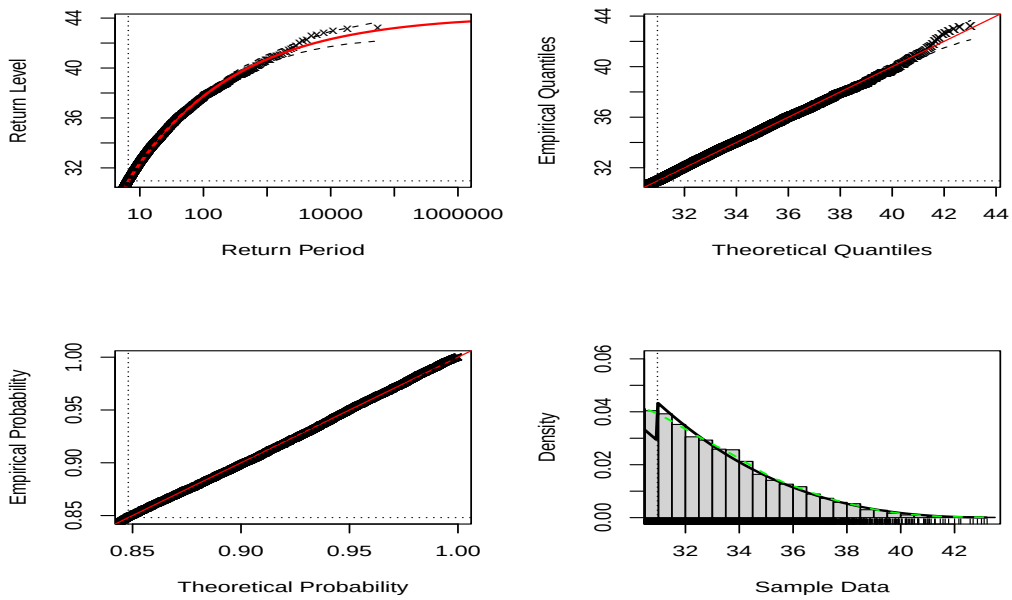

Figure 6.7: Diagnostic plots for fitted Parameterised tail fraction of Temperature.

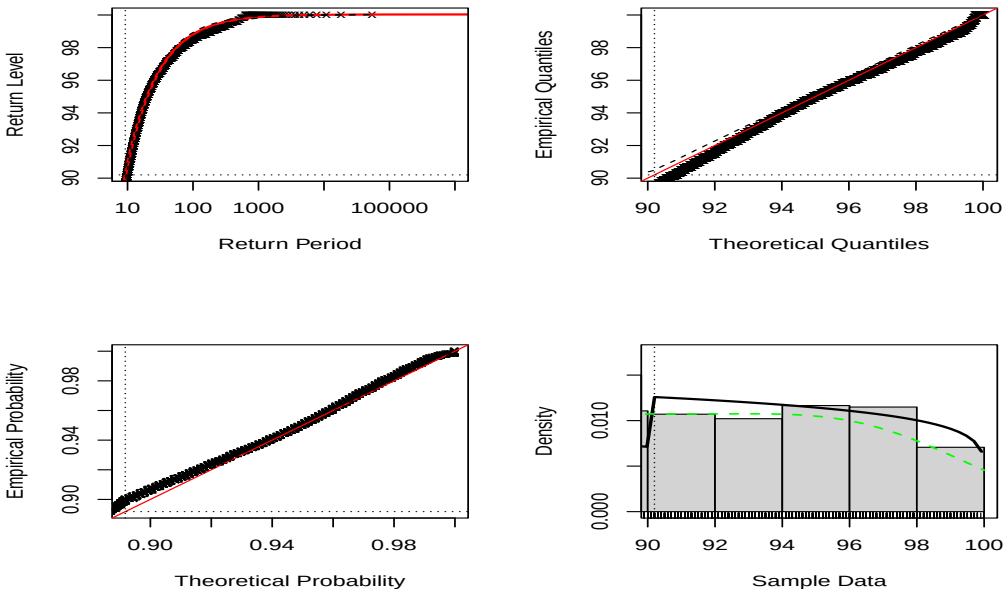

Figure 6.8: Diagnostic plots for the fitted bulk model of relative humidity.

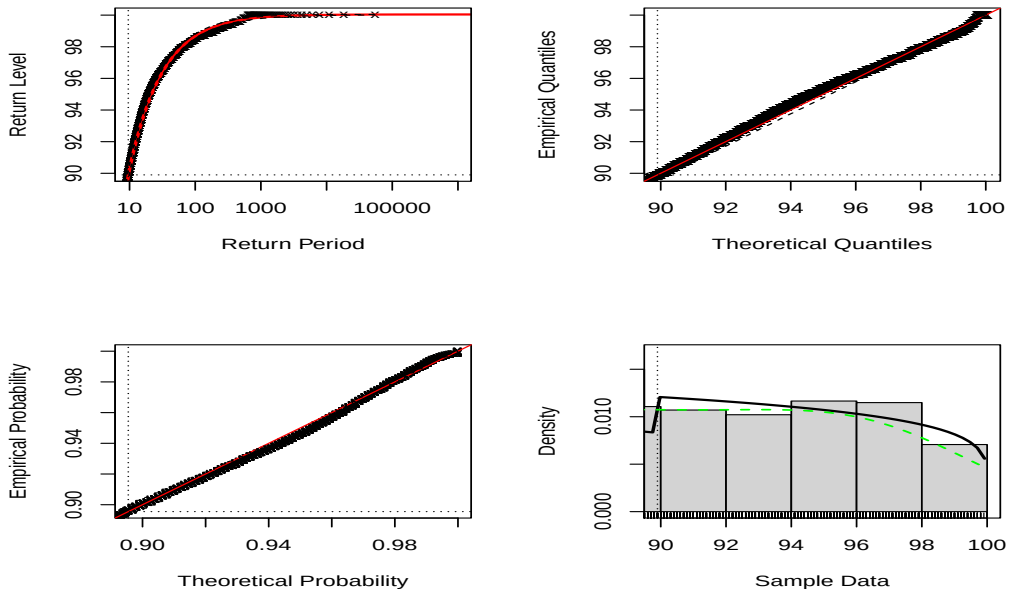

Figure 6.9: Diagnostic plots for fitted parameterised tail fraction of relative humidity

6.3.3 Quantile Estimation for Marginal Increases

Table 6.2a shows estimated quantiles based on various tail probabilities and marginal increases in GHI against Temperature. The marginal increase of the GHI converges to 0.12 W/m^2 as the temperature converges to 44.26 °C over the sampling period of 16 November 2015 to 16 November 2021. This analysis is crucial for system operators in dispatching and scheduling electrical power. In Table 6.2b, the marginal increase of the GHI converges to -0.1 W/m^2 as the RH converges to 103.26% over the sampling period of 16 November 2015 to 16 November 2021. The marginal increases of the GHI with Temperature and RH are also shown in Figure 6.10 and Figure 6.11, respectively.

Table 6.2: (**a**) Estimated quantiles at different tail probabilities and marginal increases in GHI against Temperature. (**b**) Estimated quantiles at different tail probabilities and marginal increases in GHI against RH.

(a)			
Quantile	**GHI**	**Temp (x_p)**	**Marginal Increase in GHI**
90th	1081.88	37.04	
95th	1089.08	38.24	7.20
97th	1093.70	39.01	4.62
99th	1101.62	40.33	7.92
99.9th	1112.42	42.13	10.80
99.99th	1118.30	43.11	5.88
99.999th	1121.54	43.65	3.24
99.9999th	1123.28	43.94	1.74
99.99999th	1124.24	44.10	1.50
99.999999th	1124.78	44.19	0.54
99.9999999th	1125.08	44.24	0.30
99.99999999th	1125.20	44.26	0.12
(b)			
Quantile	**GHI**	**RH (x_p)**	**Marginal Increase in GHI**
90th	421.03	96.04	
95th	414.79	97.24	−6.24
97th	410.79	98.01	−4.00
99th	403.92	99.33	−6.87
99.9th	394.56	101.13	−9.36
99.99th	389.47	102.11	−5.09
99.999th	386.66	102.65	−2.81
99.9999th	385.15	102.94	−1.51
99.99999th	384.32	103.10	−0.83
99.999999th	383.85	103.19	−0.47
99.9999999th	383.59	103.24	−0.26
99.99999999th	383.49	103.26	−0.10

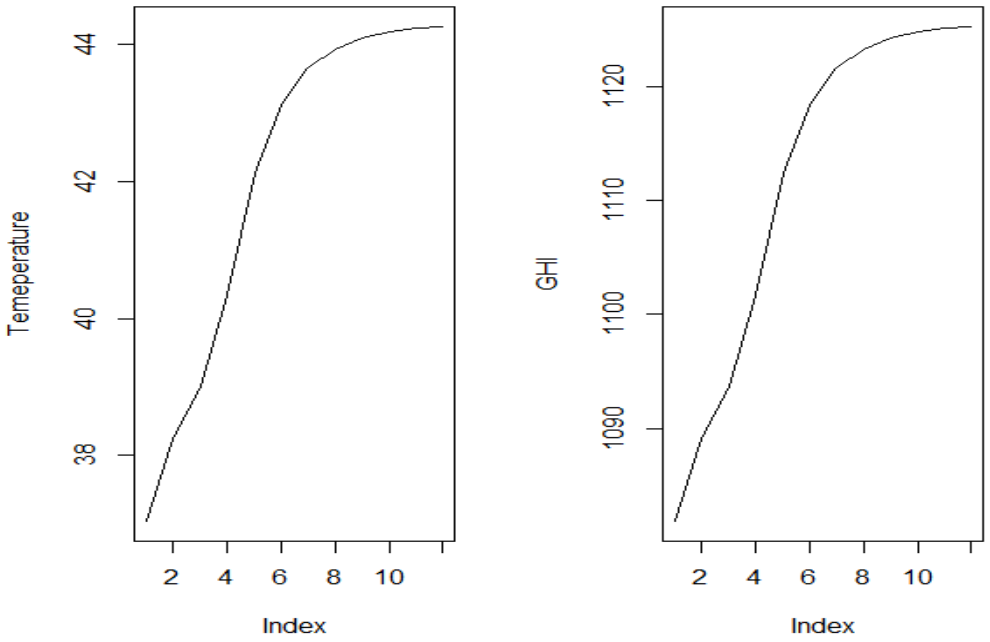

Figure 6.10: Plots for quantile estimation of temperature.

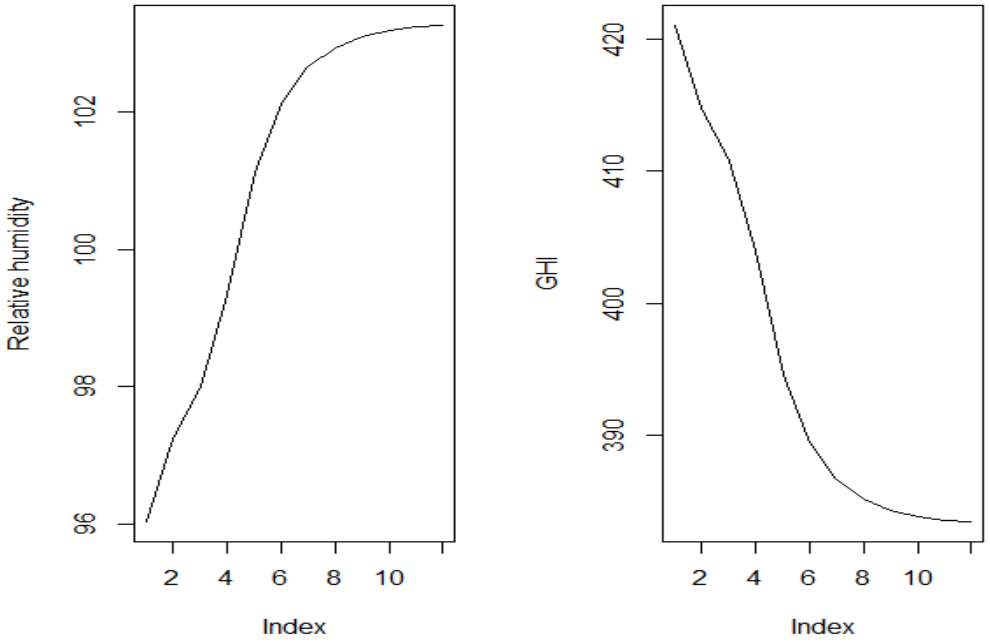

Figure 6.11: Plots for quantile estimation of RH.

6.3.4 Parameter Estimates of the Bivariate Threshold Excess Models

In Figure 6.12 for GHI, the marginal diagnostic plots show that the points are linear, indicating that the model fits the data well. Plots for probability, quantile, return level, and density for GHI data show deviations from linearity, which suggesting that GPP fit GHI data well.

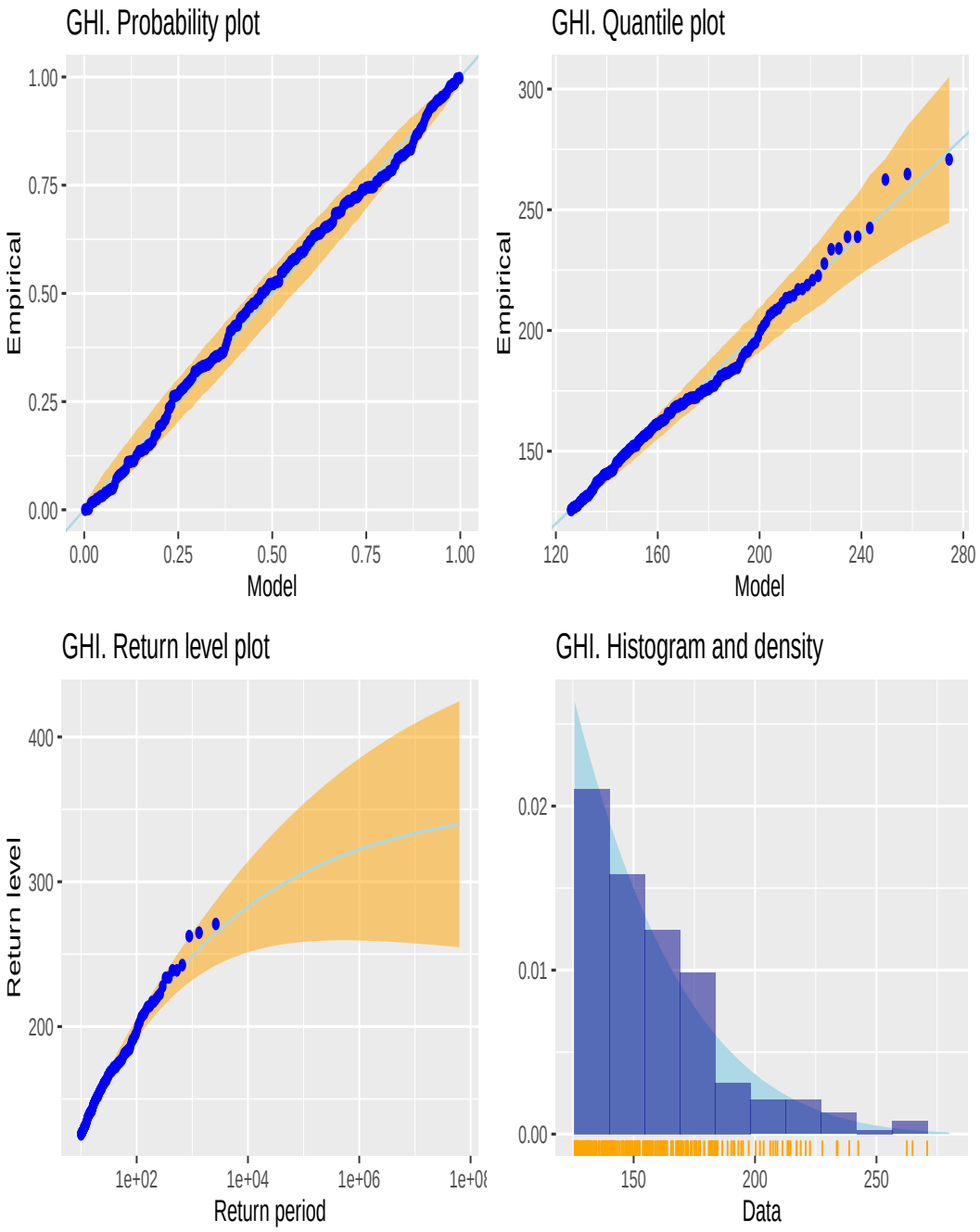

Figure 6.12: Marginal diagnostic plots for GHI.

Multivariate conditional Spearman's ρ correlation plots are shown in Figure 6.13, where ρ is Spearman's coefficient of correlation. Figure 6.13 shows a very weak positive correlation between temperature and GHI, whereas RH and GHI are negatively correlated.

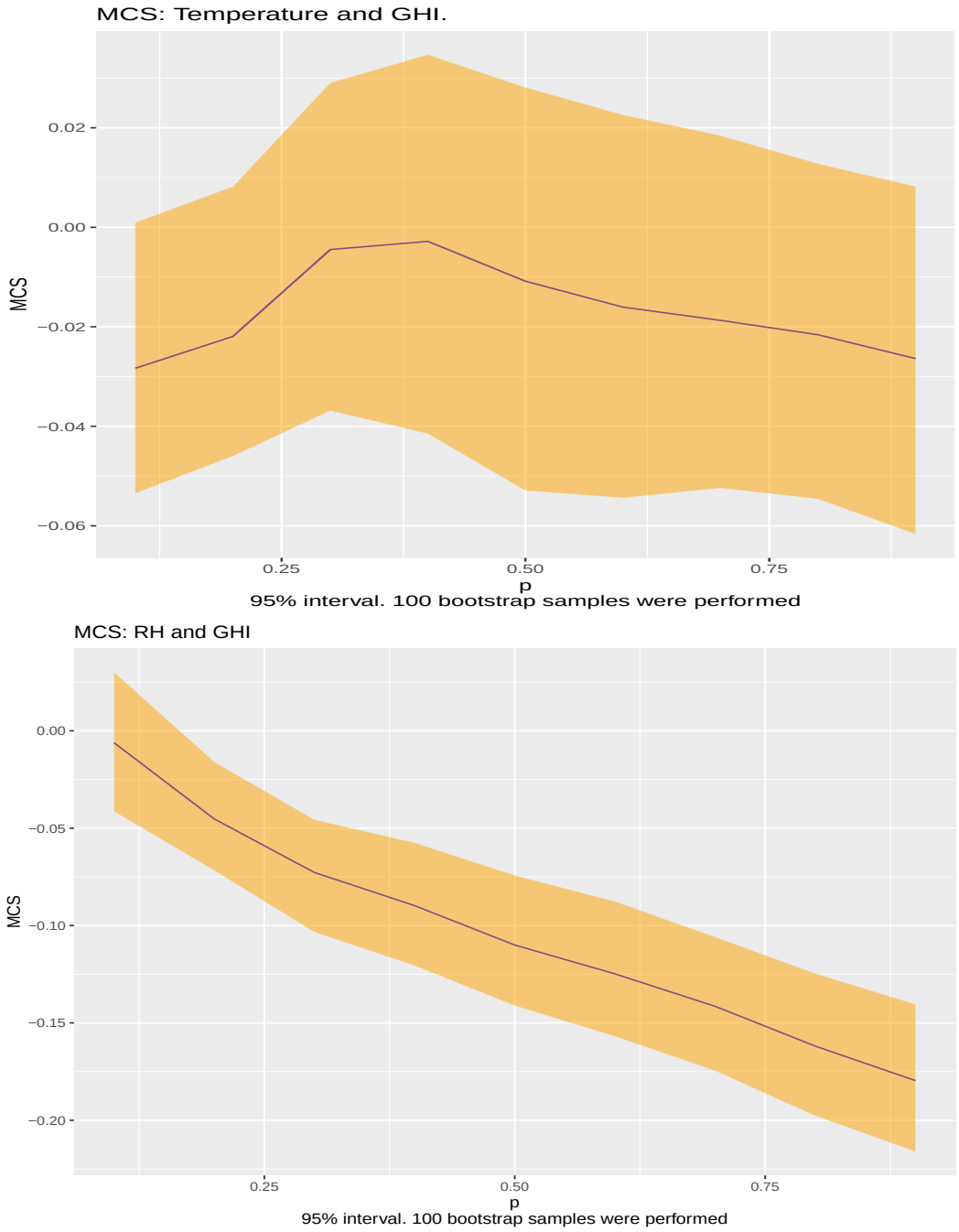

Figure 6.13: MCS plots with associated 95% bootstrap confidence intervals.

The structure of the dependence model is simplified by Laplace margins presented in Section 3.5.3. As a result, a single structure can be used to describe both positive and negative dependence (Heffernan and Tawn, 2004). A constrained Laplace margin for dependence estimation determines the threshold quantile of the transformed data as $\tau = 0.9$. Marginal quantiles defined by the threshold will be used in fitting the marginal GPD models in this study. In this study, marginal models and dependence models are used for modelling. This study focuses primarily on modelling the extremal dependence modelling of GHI on temperature and RH. Table 6.3 presents estimates of a and b, which are the target parameters of the dependence structure models. According to Heffernan and Tawn (2004), there is a strong negative and positive extremal dependence for the estimates of the parameter a close to -1 and 1. Conditioning on GHI, the estimates of the dependence parameters for temperature and RH variables are $a = -0.1187$ and $a = -0.2183$, respectively. This suggests that temperature and RH variables have a negative extremal dependence on large values of GHI.

Figure 6.14 and Figure 6.15 shows the diagnostic plots for conditioning GHI on temperature and RH, respectively. The plots, from top to bottom, display the quantiles of the fitted conditional model along with the original untransformed data, the absolute values of the scaled and centred residuals across the range of the extreme conditioning variable, and dependence model residuals across the range of the extreme conditioning variable. The horizontal lines in Figure 6.14 and Figure 6.15 are smoothest at the 90th percentile, where the parameter estimates are stable. The diagnostic plots displayed

the same characteristics for conditioning GHI on temperature and RH.

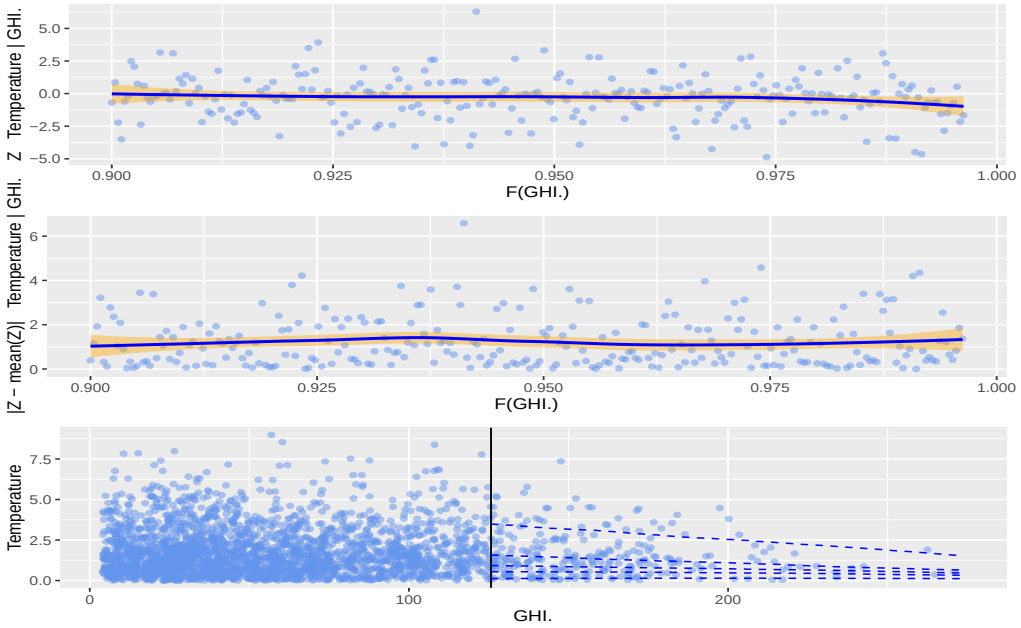

Figure 6.14: Diagnostic plots for conditioning GHI on Temperature

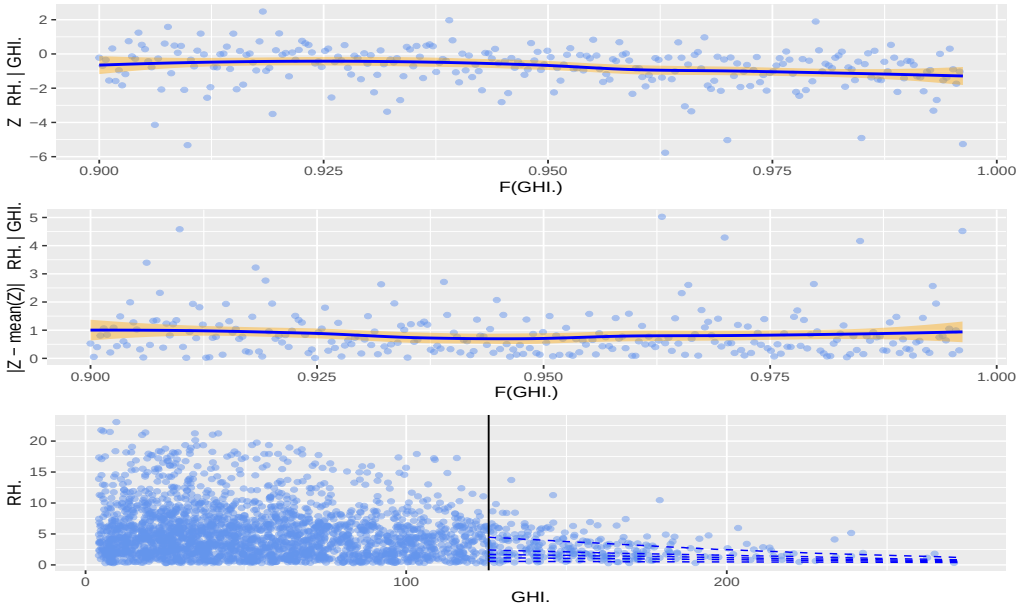

Figure 6.15: Diagnostic plots for conditioning GHI on RH.

Table 6.3: Estimates of dependence models.

	Dependence Parameters	Temperature	RH
Conditioning on GHI	a	-0.1187 *	-0.2183 *
	b	-0.2774	-0.1236

* denotes significant positive or negative values at 5% significance level.

6.3.5 Fitting Archimedean Copulas

Table 6.4 shows the estimates of bivariate Archimedean copula function parameters of the GHI against temperature and RH and the dependence structure. The estimated copula function parameters are positive for GHI against temperature. This indicates an upper-tail dependence. Based on log-likelihood, AIC and BIC, the Frank copula gives better fits because of the lowest value of AIC and BIC and the highest value of log-likelihood. The Frank copula is the only one that fits the data well for GHI against RH. The Clayton copula does not fit the data well. For the Frank copula, both upper and lower tail dependencies are zero, which means weak lower and upper tail dependencies. The parameter estimate in Table 6.4b for the Gumbel copula is 1, showing that it is at the boundary suggesting independence. As a result, a Gumbel copula is not a good fit. These results have important implications for power demand risk management.

Table 6.4: (**a**) Parameter estimation for the Archimedean copula functions of GHI against temperature. (**b**) Parameter estimation for the Archimedean copula functions of GHI against RH.

(a)

Copula	**Estimate**	ℓ	**AIC**	**BIC**	λ^U	λ^L	τ
Clayton	1.32	1717	−3431.47	−3423.26	0	0.61	0.26
Frank	4.04	5086	−10169.42	−10161.2	0	0	0.39
Gumbel	1.51	4175	−8348.91	−8340.69	1.65	0	0.34

(b)

Copula	**Estimate**	ℓ	**AIC**	**BIC**	λ^U	λ^L	τ
Clayton	0.025	−255	512	520	0.98	0	0.01
Frank	−4.10	5232	−10461.06	−10452.84	0	0	−0.40
Gumbel	1	−0.0003	2.001	10.22	1.5	0	0

The contour plots of the Archimedean copula functions are shown in Figure 6.16. As can be observed from the figures, the estimated Frank and Gumbel copulas for GHI against temperature are characterised by lower and upper tails dependence. Only the lower tail dependence characterises the estimated Clayton copula. The contour plots show that the Clayton and Gumbel copula for GHI against RH cannot fit the data effectively.

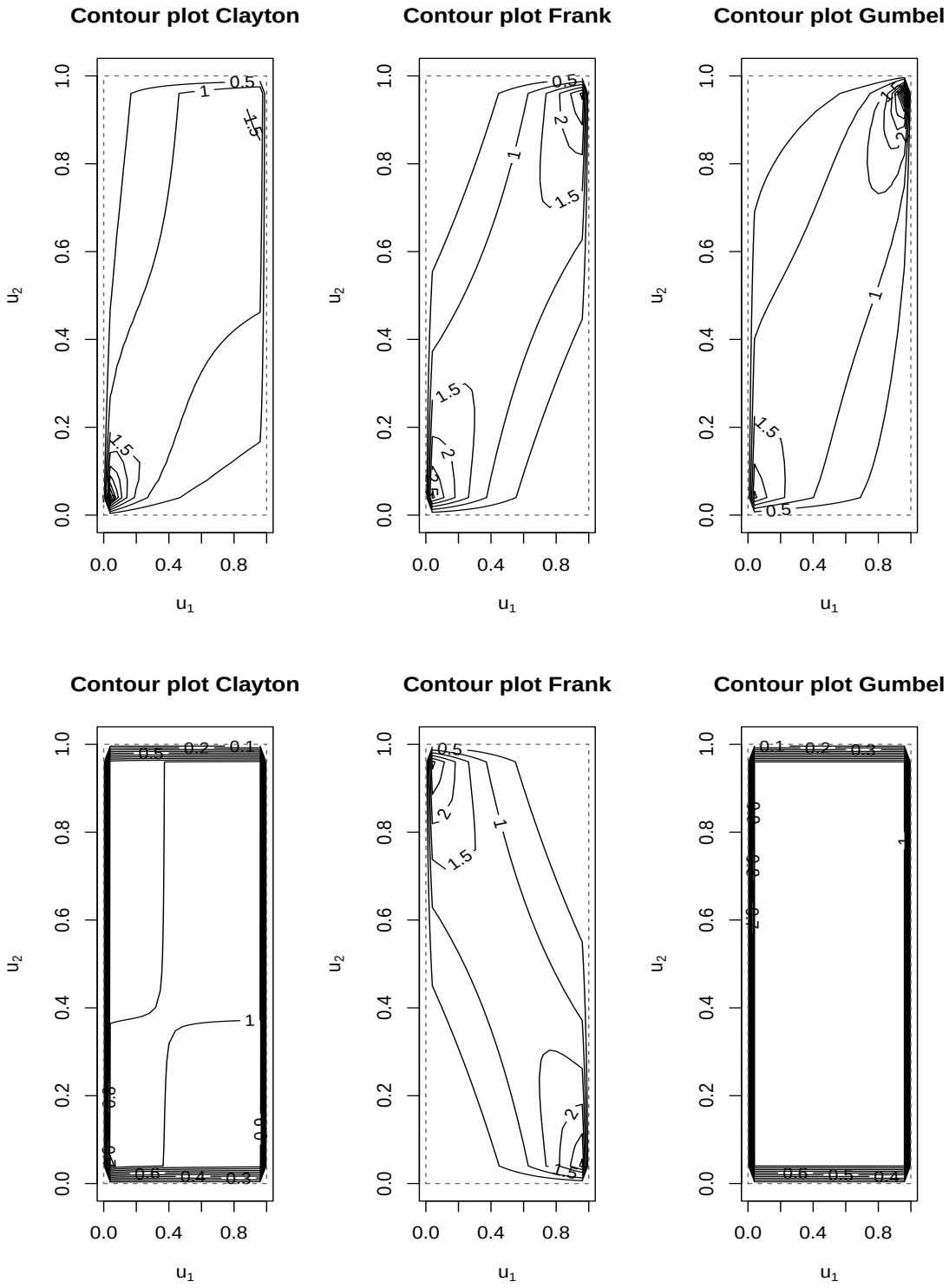

Figure 6.16: **Top panel:** Contour plots of the joint distribution for the Frank, Clayton and Gumbel for GHI against temp. **bottom panel:** Contour plots of the joint distribution for the Frank, Clayton and Gumbel for GHI against RH.

6.3.6 Mixture of Archimedean Copulas

The Frank and Gumbel copulas are used to create the mixture copula because they are the better fit for GHI against temperature from Table 6.4 and Figure 6.16 top panel. The Frank-Gumbel copula also gives a good fit to the GHI-temperature data. Table 6.5 shows the log-likelihood, AIC and BIC of the mixture of copula for GHI against temperature. Figure 6.17 shows the range of bivariate behaviour represented by a Frank-Gumbel copula. As can be observed from the figures, the estimated Frank-Gumbel copula for GHI against temperature is characterised by lower and upper tails dependence.

Table 6.5: Frank-Gumbel mixture (GHI, temperature).

Parameter	**Estimate**	**Copula**	**Std. Error**
θ_1	4.0556	Frank	0.055
θ_2	1.5726	Gumbel	0.118
w_1	0.9513	Frank	0.243
w_2	0.0487	Gumbel	N/A
AIC	−10168.58		
BIC	−10135.71		
Log-likelihood	5088.29		
λ^U	0.2087		
λ^L	0		

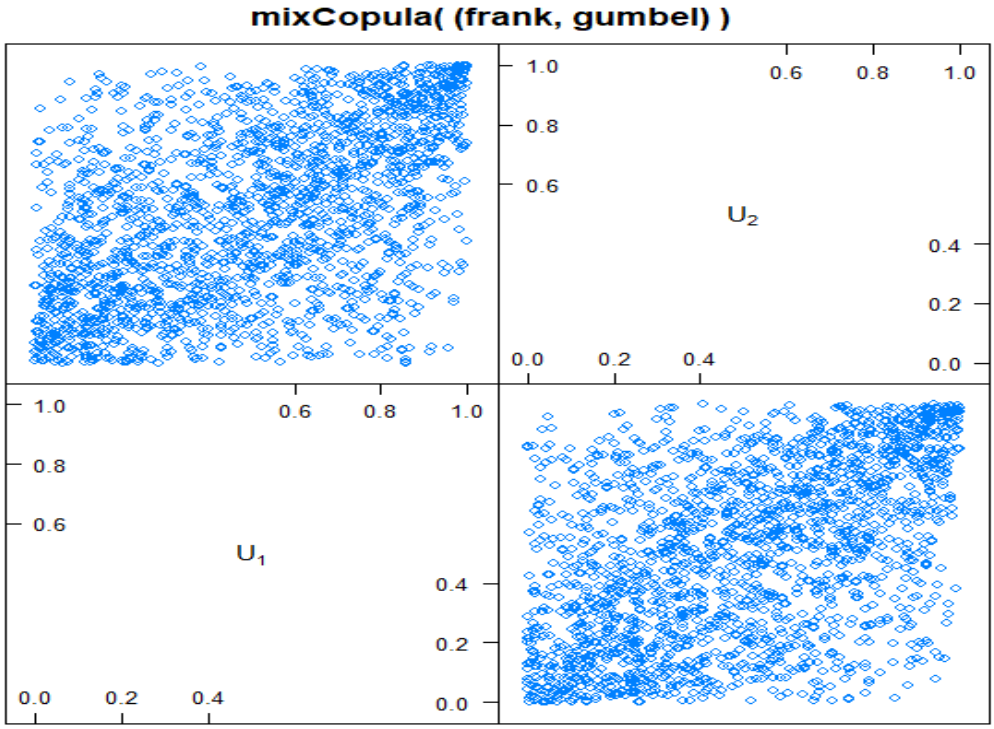

Figure 6.17: Simulated standard uniform random variables under a MixCopula for GHI against temp.

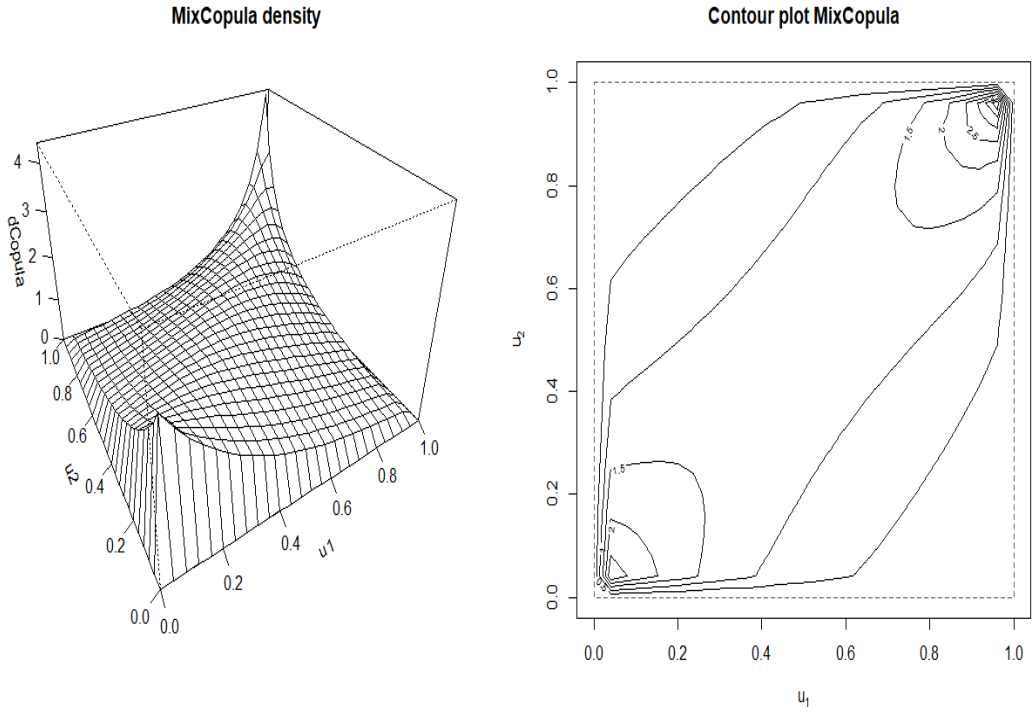

Figure 6.18: Density and Contour plots of the joint distribution for the MixCopula of GHI against temp.

6.4 Discussion

This study was conducted on one radiometric station in South Africa using data for the period 16 November 2015 to 16 November 2021 to explore the extreme dependence of GHI on temperature and RH. MARS models were used to estimate the influence of two weather variables, temperature and relative humidity, respectively, on GHI. A plot of how the GHI output relates to temperature with the best line estimated by the MARS model is found in Figure 6.2. It is shown that when the temperature increased between 0 °C and 19 °C, the GHI increased slightly and when the temperature increased between 19 °C and 30 °C, the GHI increased significantly. Again, when the temperature increased between 30 °C and 44 °C, the GHI increased slightly, which means that the relationship between temperature and GHI is directly proportional though not linear.

Figure 6.3 showed a plot of how the GHI output relates to RH with the best line estimated by the MARS model. It is shown that when RH increased between 0% and 48%, the GHI decreased slightly, and when the RH increased between 48% and 84%, the GHI decreased significantly. Again, when RH increased between 84% and 100%, the GHI decreased slightly, which means that the relationship between RH and GHI is inversely proportional. This study examined the effect of marginal temperature increases above 30 °C on GHI. When $T_t > 30$ °C, we used

$$\hat{GHI} = 420 + 24 * \max(0, T_t - 19) \tag{6.4.1}$$

As the temperature increases from 30 to 31 °C, the marginal increase in GHI is 24 W/m^2, leading to 420 + 24 W/m^2, giving us a total of 444 W/m^2. For RH, we are interested in the effect of a marginal increase in RH above 84% on GHI. When $\text{RH}_t > 84\%$, we use

$$\hat{\text{GHI}} = 540 - 12 * \max(0, \text{RH}_t - 48) \tag{6.4.2}$$

As the RH increases from 84 to 85%, the marginal increase in GHI is –12 W/m^2, leading to 540–12 W/m^2, giving us a total of 528 W/m^2. This implies that when RH increases, the GHI decreases. Estimated quantiles at various tail probabilities and marginal increases in GHI are given in Table 6.2a,b for GHI against temperature and GHI against RH, respectively. The marginal increase in GHI converges to 0.12 W/m^2 when the temperature over the sampling period from 16 November 2015 to 16 November 2021 converges to 44.26 °C. As RH converges to 103.26%, the marginal increase of GHI converges to –0.1 W/m^2.

In Section 6.3.4, the exploratory data analysis using the MCS plots was undertaken before bivariate conditional modelling of the dependence structure. The conditional bivariate extreme value modelling results are summarised in Table 6.3. The diagnostic plots for conditioning GHI on temperature and RH are presented in Figure 6.14. The dependence structure of GHI against temperature and RH was examined using the Archimedean copula functions such as Clayton, Frank, and Gumbel. The Frank copula for GHI against temperature better fits the data based on log-likelihood, AIC, and BIC since it has the lowest AIC and BIC values and the highest log-likelihood. For GHI against RH, only the Frank copula adequately fits the data. The con-

tour plots of the Archimedean copula functions were presented in Figure 6.16.

To our knowledge, no work has been done using conditional multivariate extreme value modelling of the extremal dependence modelling of GHI with the temperature and GHI with relative humidity. The use of copula models is discussed in the literature. From the study done by Yet and Masseran (2021), the Gaussian copula, which is in the elliptical family, was found to be the best fitting to the clearness index and solar radiation data. The Clayton copula, which belongs to the Archimedean family, yielded the best performance based on the work done by Bazrafshan et al. (2015) using the mean monthly solar radiation and sunshine duration hours data. Based on our studies, the Frank copula best fits GHI and temperature data, including the GHI and relative humidity data. The extremal dependence modelling from this study will help the decision-makers and planners in the energy sector plan the effects of extreme temperature and RH on GHI outputs.

6.5 Conclusion

This study presented an extremal dependence modelling of GHI with temperature and RH at one radiometric station using South African data from 16 November 2015 to 16 November 2021. In comparison, some studies have been conducted on the modelling of GHI, as well as a few studies on modelling GHI in South Africa (see Chandiwana et al. (2021); Mutavhatsindi et al. (2020)), no evidence of studies have been conducted on modelling extremal dependence of GHI with temperature and RH using the MARS, extreme value theory and copula models. This is the gap this study bridged by modelling

the effect of temperature and RH on GHI and the extremal dependence of GHI on temperature and RH. The results from the MARS models revealed that the relationship between GHI and temperature is direct, whereas the relationship between GHI and RH is inverse proportionality.

The marginal increase at different quantiles also confirmed that when the temperature increases, the GHI also increases, and when the RH increases, the GHI decreases. The conditional multivariate extreme value theory was used in modelling the dependence features on GHI with temperature and RH. Temperature and RH marginal diagnostic plots are linear, indicating that the parameterised tail fraction is appropriate for the data. In MCS plots, temperature and GHI are weakly correlated, but RH and GHI are negatively correlated. Conditioning on GHI helps to understand the significant negative extremal dependence of temperature and RH on large values of GHI. One of our findings is that the dependence structure between GHI and variable temperature and RH is asymmetric. Furthermore, the Frank copula is the best-fitting model for variable temperature and RH, implying the presence of extreme co-movements. The results from this study could be useful to decision-makers in power utility companies that face uncertainty in GHI power production due to extreme temperatures and RH.

Chapter 7

Conclusion

7.1 Introduction

Electricity demand forecasting is essential for the operation of power utilities. During load flow analysis, scheduling of start-up times for peak stations, determining consistent and reliable supply schedules and determining optimal distribution of electrical energy, modelling extreme peak electricity demand and extreme peak global horizontal irradiance (GHI) power production forecasting is very important to system operators and decision-makers in the electricity sector. In this thesis, we modelled and analysed statistical models for forecasting extreme electricity demand during peak hours, the probabilistic forecasting of hourly GHI at extremely high quantiles and the extremal dependence modelling of GHI with temperature and relative humidity (RH) in South Africa. This chapter is presented as follows. Sections 7.2 and 7.3, respectively, present the overall summary of modelling and key findings. The study's concluding remarks are discussed in Section 7.4. Limitations of the study and areas of future research directions are given in Sections 7.5 and 7.6, respectively.

7.2 Summary of modelling by chapters

In Chapter 4, this study presents an application of linear quantile regression (LQR), non-linear quantile regression (NLQR), and additive quantile regression (AQR) models for forecasting extreme electricity demand at peak hours which are 18:00, 19:00, 20:00 and 21:00 using the Northern Cape energy data. Variables are selected based on the least absolute shrinkage and selection operator. The forecast evaluation of the proposed models is done using the pinball loss function (PLF). This chapter uses a mixed-integer programming approach with priority constraints to solve the unit commitment problem.

In Chapter 5, we discuss comparisons of a semi-parametric extremal mixture (SPEM), generalised additive extreme value (GAEV) or quantile regression via asymmetric Laplace distribution (QR-ALD), additive quantile regression (AQR-1), additive quantile regression with temperature variable (AQR-2) and penalised cubic regression smoothing spline (benchmark) models for forecasting hourly global horizontal irradiance (GHI) at extreme quantiles. With the continuous ranked probability score (CRPS) and the PLF, the forecast evaluation of the proposed models is carried out. Chapter 6 discusses extremal dependence modelling of GHI with temperature and relative humidity (RH). Multivariate adaptive regression splines (MARS) models, extreme value theory, and copula models are employed in this chapter to determine the maximum temperature at which maximum GHI will be generated.

7.3 Modelling discussion and summary of key findings

The modelling discussion and summary of key findings considered in this study are as follows:

1 AQR is the best-fitting for hours 18:00 and 19:00, whereas LQR is best-fitting for hours 20:00 and 21:00. A mixed-integer programming approach is used to solve the unit commitment problem in this chapter. Based on the unit commitment problem results, all the electric generation sources, including hydroelectric, wind, concentrated solar, and solar photovoltaic, are less costly.

2 The AQR-2 is the most accurate model for quantiles ($\tau = 0.95$, 0.97, 0.99 and 0.999), and the GAEV model is the most accurate model for quantile ($\tau = 0.9999$).

3 Using the MARS models, two thresholds were estimated as 19^{o}C and 30^{o}C, respectively. For temperature values, less than 19^{o}C temperature was found to increase slightly at a 17 W/m^2 per degree increase, for temperature values between 19^{o}C and 30^{o}C GHI increases significantly at the rate of 41 W/m^2 per degree increase in temperature. Lastly, for temperature values above 30^{o}C, the rate of increase was 24 W/m^2 per degree increase. On applying extreme value theory models, the marginal increases of GHI converged to 0.12 W/m^2 as temperature converged to 44.26^{o}C. Similarly, for GHI versus RH, the two identified thresholds were 48% and 84%, respectively. For RH less than 48%, GHI decreased by 1.1%, while between 48% and 84%, it decreased by

12% and by 5% for RH above 84%. Using MRAS coupled with the GPD model, the marginal increase of GHI converged to -0.1 W/m^2 as RH converged to 103.26%, a sign of over saturation of the air.

(4) Using conditional multivariate extreme value theory to model the dependence, it was established that temperature and GHI are weakly correlated, but RH and GHI are negatively correlated. The dependence structure between GHI and temperature and GHI and RH is asymmetric. Furthermore, the Frank copula is the best-fitting model for variable temperature and RH, implying the presence of extreme co-movements.

7.4 Concluding remarks

Due to the growing, volatile and unexpected nature of the energy supply, demand, and prices, decision-makers in the industry need precise forecasting for a future market that is unclear. Power utility operators might benefit from the models discussed in this study for optimally integrating highly intermittent renewable energies into the grid. In light of the many modelling approaches suggested and significant findings, the study offers the following responses to the objectives listed in Chapter 1: First, models for peak electricity demand forecasting at the extreme quantile are proposed, including LQR, NLQR, and AQR models. Using the PLF, the predicting accuracy of the LQR, NLQR, and AQR models is compared. When employing mixed-integer programming to solve the unit commitment problem, out-of-sample forecasts are used. The development of models for predicting hourly extreme

peak electricity demand is the key contribution of this study. These findings could be helpful to system operators in the energy sector who must schedule and dispatch electricity at peak times when the grid is constrained owing to peak load demand to maintain the lowest cost.

Second, for probabilistic forecasting of hourly GHI at extremely high quantiles, the study conducted a comparative analysis of SPEM, GAEV or QR-ALD, AQR-1, AQR-2 and penalised cubic regression smoothing spline (benchmark) models. At various quantile levels, the comparative analysis is conducted. Using PLF and CRPS, the predicting accuracy of the SPEM, GAEV or QR-ALD, AQR-1, AQR-2, and Benchmark models is compared. Lastly, the study also employs MARS, extreme value theory and copula models to examine the extreme dependence modelling of GHI with temperature and RH.

7.5 Limitations of the thesis

In this thesis, we modelled the extreme electricity demand at peak hours using Northern Cape data. It would have been interesting to analyse the extreme electricity demand for South Africa and each of the country's nine provinces. The study used mixed-integer programming to solve the unit commitment problem. However, it would be good to use other optimal solution approaches for comparative analysis. The study also has limitations due to the exclusion of cloud cover as a covariate. Despite using covariates such as temperature and RH, including cloud cover could improve the results when modelling GHI.

7.6 Future research studies

The results of this study provide some possible areas for future research as follows:

1 Future research could look at spatio-temporal modelling of extreme global horizontal irradiance, including other forms of renewable energies such as wind power.

2 This study was limited by the absence of cloud cover as a covariate which will be addressed in future studies.

3 In this thesis, Archimedean copulas were used in modelling the extremal dependence of global horizontal irradiance with air temperature. Future research could explore the application of bivariate vine copulas, which are known to be flexible in high-dimensional dependence modelling.

4 Furthermore, it will be more interesting to use multi-objective programming to solve the unit commitment problem in which one objective would be to minimise the total generation cost and the other to minimise greenhouse gas emissions.

5 It is possible to use a full Bayesian MARS model in the future, in which the number of splines, knot points, and all other parameters are treated as unknowns.

References

A.A. Adeala, Z. Huan, and C.C. Enweremadu. Evaluation of global solar radiation using multiple weather parameters as predictors for south africa provinces. *Thermal Science*, 19(suppl. 2):495–509, 2015.

J. Aghaei, A. Ahmadi, H.A. Shayanfar, and A. Rabiee. Mixed integer programming of generalized hydro-thermal self-scheduling of generating units. *Electrical Engineering*, 95(2):109–125, 2013.

F.N. Al Farsi, M.H. Albadi, N. Hosseinzadeh, and A.H/ Al Badi. Economic dispatch in power systems. In *2015 IEEE 8th GCC Conference & Exhibition*, pages 1–6. IEEE, 2015.

S. Alessandrini, L. Delle Monache, S. Sperati, and G. Cervone. An analog ensemble for short-term probabilistic solar power forecast. *Applied energy*, 157:95–110, 2015.

A.A. Alresheedi and M.A. Al-Hagery. Forecasting the global horizontal irradiance based on boruta algorithm and artificial neural networks using a lower cost. *International Journal of Advanced Computer Science and Applications*, 11(9):79–92, 2020.

J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F.J Martinez-de Pison, and F. Antonanzas-Torres. Review of photovoltaic power forecasting. *Solar Energy*, 136:78–111, 2016.

V. Arakelian and D. Karlis. Clustering dependencies via mixtures of copulas. *Communications in Statistics-Simulation and Computation*, 43(7):1644–1661, 2014.

E. Arriagada, E. Lopez, M. López, R. Blasco-Gimenez, C. Roa, and M. Poloujadoff. A probabilistic economic dispatch model and methodology considering renewable energy, demand and generator uncertainties. *Electric Power Systems Research*, 121:325–332, 2015.

R. Azimi, M. Ghayekhloo, and M. Ghofrani. A hybrid method based on a new clustering technique and multilayer perceptron neural networks for hourly solar radiation forecasting. *Energy Conversion and Management*, 118:331–344, 2016.

J. Bazrafshan, N. Heidari, I. Moradi, and Z. Aghashariatmadary. Simultaneous stochastic simulation of monthly mean daily global solar radiation and sunshine duration hours using copulas. *Journal of Hydrologic Engineering*, 20(4):04014061, 2015.

C.N. Behrens, H.F. Lopes, and D. Gamerman. Bayesian analysis of extreme events with threshold estimation. *Statistical modelling*, 4(3):227–244, 2004.

J. Beirlant, Y. Goegebeur, J. Segers, and J.L. Teugels. *Statistics of Extremes: Theory and Applications*, volume 558. John Wiley & Sons, 2004.

J. Beirlant, Y. Goegebeur, J. Segers, and J.L. Teugels. *Statistics of extremes: theory and applications*. John Wiley & Sons, 2006.

F. Cavallaro. Electric load analysis using an artificial neural network. *International journal of energy research*, 29(5):377–392, 2005.

E. Chandiwana, C. Sigauke, and A. Bere. Twenty-four-hour ahead probabilistic global horizontal irradiance forecasting using gaussian process regression. *Algorithms*, 14(6):177, 2021.

D.K. Chaturvedi and I. Isha. Solar power forecasting: A review. *International Journal of Computer Applications*, 145(6):28–50, 2016.

D.G. Clayton. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1):141–151, 1978.

S. Coles. An introduction to statistical modeling of extreme values, 2001.

M. Conde Amboage, W. González Manteiga, and C. Sánchez Sellero. Quantile regression: estimation and lack-of-fit tests. *BEIO, Boletín de Estadística e Investigación Operativa*, 34(2):97–116, 2018.

S. Corbella and D.D. Stretch. Simulating a multivariate sea storm using archimedean copulas. *Coastal Engineering*, 76:68–78, 2013.

P. Craven and G. Wahba. Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross–validation. *Numerische Mathematik*, 31:317–403, 1979.

H. Dai, N. Zhang, and W. Su. A literature review of stochastic programming and unit commitment. *Journal of Power and Energy Engineering*, 3(04): 206, 2015.

U.K. Das, K.S. Tey, M. Seyedmahmoudian, S. Mekhilef, M.Y.I. Idris, W. Van Deventer, B. Horan, and A. Stojcevski. Forecasting of photovoltaic power generation and model optimization: A review. *Renewable and Sustainable Energy Reviews*, 81:912–928, 2018.

C. Davino, M. Furno, and D. Vistocco. *Quantile regression: theory and applications*, volume 988. John Wiley & Sons, 2013.

A.P. Dawid and P. Sebastiani. Coherent dispersion criteria for optimal experimental design. *Annals of Statistics*, pages 65–81, 1999.

D.K. Dey and J. Yan. *Extreme value modeling and risk analysis: methods and applications*. CRC Press, 2016.

M. Diagne, M. David, P. Lauret, J. Boland, and N. Schmutz. Review of solar irradiance forecasting methods and a proposition for small-scale insular grids. *Renewable and Sustainable Energy Reviews*, 27:65–76, 2013.

G. Durrieu, I. Grama, Q. Pham, and J. Tricot. Nonparametric adaptive estimation of conditional probabilities of rare events and extreme quantiles. *Extremes*, 18(3):437–478, 2015.

G. Durrieu, I. Grama, K. Jaunatre, Q. Pham, and J. Tricot. extremefit: An r package for extreme quantiles. 2016.

N. Elamin. Quantile regression model for peak load demand forecasting with approximation by triangular distribution to avoid blackouts. 2018.

S. Fan and R.J. Hyndman. Short-term load forecasting based on a semi-parametric additive model. *IEEE Transactions on Power Systems*, 27(1): 134–141, 2012.

M. Fasiolo, Y. Goude, R. Nedellec, and S.N. Wood. Fast calibrated additive quantile regression. *arXiv preprint arXiv:1707.03307*, 2017.

M. Fasiolo, S.N. Wood, Y. Goude, R. Nedellec, and M. Zaffran. Fast calibrated additive quantile regression. *Journal of the American Statistical Association*, pages 1–11, 2020a.

M. Fasiolo, S.N. Wood, M. Zaffran, R. Nedellec, and Y. Goude. qgam: Bayesian non-parametric quantile regression modelling in r. *arXiv preprint arXiv:2007.03303*, 2020b.

E.A. Feinberg and D. Genethliou. Load forecasting. In *Applied mathematics for restructured electric power systems*, pages 269–285. Springer, 2005.

L.A. Fernandez-Jimenez, A. Muñoz-Jimenez, A. Falces, M. Mendoza-Villena, E. Garcia-Garrido, P.M. Lara-Santillan, E. Zorzano-Alba, and P.J. Zorzano-Santamaria. Short-term power forecasting system for photovoltaic plants. *Renewable Energy*, 44:311–317, 2012.

R.A. Fisher and L.H.C. Tippett. Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 24, pages 180–190. Cambridge University Press, 1928.

M.J. Frank. On the simultaneous associativity off (x, y) andx+ y- f (x, y). *Aequationes mathematicae*, 19(1):194–226, 1979.

P. Friederichs and T.L. Thorarinsdottir. Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction. *Environmetrics*, 23(7):579–594, 2012.

J. Friedman. Invited paper multivariate adaptive regression splines '. *The Annals of Statistics*, 19(1):1–141, 1991.

S. Fukutome, M.A. Liniger, and M. Süveges. Automatic threshold and run parameter selection: a climatology for extreme hourly precipitation in switzerland. *Theoretical and Applied Climatology*, 120(3):403–416, 2015.

Y. Fukuyama and H. Yoshida. A particle swarm optimization for reactive power and voltage control in electric power systems. In *Proceedings of the 2001 Congress on Evolutionary Computation (IEEE Cat. No. 01TH8546)*, volume 1, pages 87–93. IEEE, 2001.

A. Gaba, I. Tsetlin, and R.L. Winkler. Combining interval forecasts. *Decision Analysis*, 14(1):1–20, 2017.

P. Gaillard, Y. Goude, and R. Nedellec. Additive models and robust aggregation for gefcom2014 probabilistic electric load and electricity price forecasting. *International Journal of forecasting*, 32(3):1038–1050, 2016.

B. Gao, X. Huang, J. Shi, Y. Tai, and J. Zhang. Hourly forecasting of solar irradiance based on ceemdan and multi-strategy cnn-lstm neural networks. *Renewable Energy*, 162:1665–1683, 2020.

B.A. Garro, H. Sossa, and R.A. Vazquez. Design of artificial neural networks using a modified particle swarm optimization algorithm. In *2009 International Joint Conference on Neural Networks*, pages 938–945. IEEE, 2009.

R. Gençay and F. Selçuk. Extreme value theory and value-at-risk: Relative performance in emerging markets. *International Journal of Forecasting*, 20(2):287–303, 2004.

M. Ghayekhloo, M. Ghofrani, M.B. Menhaj, and R. Azimi. A novel clustering approach for short-term solar radiation forecasting. *Solar Energy*, 122: 1371–1383, 2015.

M. Ghofrani and M. Alolayan. *Time series and renewable energy forecasting*, volume 10. IntechOpen, 2018.

C. Gonçalves, L. Cavalcante, M. Brito, R.J. Bessa, and J. Gama. Forecasting conditional extreme quantiles for wind energy. *Electric Power Systems Research*, 190:106636, 2021.

Y. Goude, R. Nedellec, and N. Kong. Local short and middle term electricity load forecasting with semi-parametric additive models. *IEEE transactions on smart grid*, 5(1):440–446, 2013.

E.J. Gumbel. Distributions des valeurs extremes en plusiers dimensions. *Publ. Inst. Statist. Univ. Paris*, 9:171–173, 1960.

N. Gupta and I. Ali. *Optimization with LINGO-18: Problems and Applications*. CRC Press, 2021.

A. Gxasheka, H. Mokhonoana, and R. Mokoena. Monitoring renewable energy performance of power plants. *Tracking progress of 2017 (Issue 11) March 2018 by NERSA*, 2018.

T.J. Hastie and R.J. Tibshirani. Generalized additive models, volume 43: Crc press, 1990.

J.E. Heffernan and J.A. Tawn. A conditional approach for multivariate extreme values (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):497–546, 2004.

D. Heinemann, E. Lorenz, and M. Girodo. Forecasting of solar radiation. *Solar energy resource management for electricity generation from local level to global scale. Nova Science Publishers, New York*, pages 83–94, 2006.

T. Hong and S. Fan. Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting*, 32(3):914–938, 2016.

T. Hong, P. Pinson, S. Fan, H. Zareipour, A. Troccoli, and R.J. Hyndman. Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond, 2016.

C. Hor, S. Watson, D. Infield, and S. Majithia. Assessing load forecast uncertainty using extreme value theory. *16th PSCC, Glasgow*, 2008.

L. Hu. Dependence patterns across financial markets: a mixed copula approach. *Applied financial economics*, 16(10):717–729, 2006.

Y. Hu. Extreme value mixture modelling with simulation study and applications in finance and insurance. 2013.

Y. Hu and C. Scarrott. evmix: An r package for extreme value mixture modeling, threshold estimation and boundary corrected kernel density estimation. 2018.

C. Hwang and J. Shim. A simple quantile regression via support vector machine. In *International Conference on Natural Computation*, pages 512–520. Springer, 2005.

R.J. Hyndman and S. Fan. Density forecasting for long-term peak electricity demand. *IEEE Transactions on Power Systems*, 25(2):1142–1153, 2009.

R.H. Inman, H.T.C. Pedro, and C.F.M. Coimbra. Solar forecasting methods for renewable energy integration. *Progress in energy and combustion science*, 39(6):535–576, 2013.

C. Keef, I. Papastathopoulos, and J.A. Tawn. Estimation of the conditional distribution of a multivariate variable given that one of its components is large: Additional constraints for the heffernan and tawn model. *Journal of Multivariate Analysis*, 115:396–404, 2013.

J. Kennedy and R. Eberhart. Particle swarm optimization. In *Proceedings of ICNN'95-international conference on neural networks*, volume 4, pages 1942–1948. IEEE, 1995.

Y. Kim, H.G. Son, and S. Kim. Short term electricity load forecasting for institutional buildings. energy reports, 5, 1270–1280, 2019.

R. Koenker. *Quantile Regression*. Cambridge University Press, 2005. URL https://EconPapers.repec.org/RePEc:cup:cbooks:9780521608275.

R. Koenker. Quantile regression: 40 years on. *Annual review of economics*, 9:155–176, 2017.

R. Koenker and G.W. Bassett. Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50, 1978.

R. Koenker, S. Portnoy, P.T. Ng, A. Zeileis, P. Grosjean, and B.D. Ripley. Package 'quantreg'. *Cran R-project. org*, 2018.

V. Kostylev, A. Pavlovski, et al. Solar power forecasting performance–towards industry standards. In *1st international workshop on the integration of solar power into power systems, Aarhus, Denmark*. Energynautics GmbH Mühlstraße Langen, Germany, 2011.

M. Kurban and U.B. Filik. Unit commitment scheduling by using the autoregressive and artificial neural network models based short-term load forecasting. In *Proceedings of the 10th International Conference on Probablistic Methods Applied to Power Systems*, pages 1–5. IEEE, 2008.

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*, 173:566–577, 2018.

M.E. Lebotsa, C. Sigauke, A. Bere, R. Fildes, and J.E. Boylan. Short term electricity demand forecasting using partially linear additive quantile regression with an application to the unit commitment problem. *Applied Energy*, 222:104–118, 2018.

D.H.W. Li, W. Chen, S. Li, and S. Lou. Estimation of hourly global solar radiation using multivariate adaptive regression spline (mars)–a case study of hong kong. *Energy*, 186:115857, 2019.

Y. Li, Y. He, Y. Su, and L. Shu. Forecasting the daily power output of a grid-connected photovoltaic system based on multivariate adaptive regression splines. *Applied Energy*, 180:392–401, 2016.

C. Lyons and C. Gross. Power generation technology data for integrated resource plan of south africa. 2015.

A. MacDonald, C.J. Scarrott, D. Lee, B. Darlow, M. Reale, and G. Russell. A flexible extreme value mixture model. *Computational Statistics & Data Analysis*, 55(6):2137–2157, 2011.

D. Maposa, A.M. Seimela, C. Sigauke, and J.J. Cochran. Modelling temperature extremes in the limpopo province: Bivariate time-varying threshold excess approach. *Natural Hazards*, 107(3):2227–2246, 2021.

N. Maswanganyi, C. Sigauke, and E. Ranganai. Peak electricity demand forecasting using partially linear additive quantile regression models. In *Annual Proceedings of the South African Statistical Association Conference*, volume 2017, pages 25–32. South African Statistical Association (SASA), 2017.

N. Maswanganyi, C. Sigauke, and E. Ranganai. Prediction of extreme conditional quantiles of electricity demand: An application using south african data. *Energies*, 14(20):6704, 2021.

S. Milborrow. earth: Multivariate adaptive regression splines. r package version 4.6. 3, 2018.

M. Mishra, P.B. Dash, J. Nayak, B. Naik, and S.K. Swain. Deep learning and wavelet transform integrated approach for short-term solar pv power prediction. *Measurement*, 166:108250, 2020.

M. Mohandes. Support vector machines for short-term electrical load forecasting. *International Journal of Energy Research*, 26(4):335–345, 2002.

P. Mpfumali, C. Sigauke, A. Bere, and S. Mulaudzi. Day ahead hourly global horizontal irradiance forecasting—application to south african data. *Energies*, 12(18):3569, 2019.

J. Munkhammar and J. Widén. Correlation modeling of instantaneous solar irradiance with applications to solar engineering. *Solar Energy*, 133:14–23, 2016.

N.C. Mupondo, C.B. Jones, and C. Sigauke. The volatility spillovers between zimbabwe, the united states of america, south africa, botswana and china: Copula garch model. pages 105–139, 2021.

T. Mutavhatsindi, C. Sigauke, and R. Mbuvha. Forecasting hourly global horizontal solar irradiance in south africa using machine learning models. *IEEE Access*, 8:198872–198885, 2020.

R. Nageem and R. Jayabarathi. Predicting the power output of a grid-connected solar panel using multi-input support vector regression. *Procedia computer science*, 115:723–730, 2017.

N. Naifar. Modelling dependence structure with archimedean copulas and applications to the itraxx cds index. *Journal of Computational and Applied Mathematics*, 235(8):2459–2466, 2011.

S. Nam and J. Hur. Probabilistic forecasting model of solar power outputs based on the naïve bayes classifier and kriging models. *Energies*, 11(11): 2982, 2018.

R.B. Nelson. An introduction to copulas, 2006.

M.M. Nemukula, C. Sigauke, and D. Maposa. Bivariate threshold excess models with application to extreme high temperatures in limpopo province of south africa. In *Annual Proceedings of the South African Statistical Association Conference*, volume 2018, pages 33–40. South African Statistical Association (SASA), 2018.

H.B. Olaiya. Transforming our world: The 2030 agenda for sustainable development & international decade for people of african descent. 2016.

G.J. Osório, J.M. Lujano-Rojas, J.C.O. Matias, and J.P.S. Catalão. A probabilistic approach to solve the economic dispatch problem with intermittent renewable energy sources. *Energy*, 82:949–959, 2015.

A. Pierrot and Y. Goude. Short-term electricity load forecasting with generalized additive models. *Proceedings of ISAP power*, 2011:593–600, 2011.

A. Pillay, R. Stephen, and H. Geldenhuys. Solar photovoltaic energy output forecasting for the south african power system. *Energy*, 18:20, 2014.

J.I. Prieto, J.C. Martínez-García, and D. García. Correlation between global solar irradiation and air temperature in asturias, spain. *Solar Energy*, 83 (7):1076–1085, 2009.

E. Ranganai and C. Sigauke. Capturing long-range dependence and harmonic phenomena in 24-hour solar irradiance forecasting: A quantile regression robustification via forecasts combination approach. *IEEE Access*, 8:172204–172218, 2020.

M.Q. Raza, M. Nadarajah, and C. Ekanayake. On recent advances in pv output power forecast. *Solar Energy*, 136:125–144, 2016.

B. Saravanan, S. Das, S. Sikri, and D.P. Kothari. A solution to the unit commitment problem—a review. *Frontiers in Energy*, 7(2):223–236, 2013.

C. Scarrott and A. MacDonald. A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT-Statistical journal*, 10(1): 33–60, 2012.

B.M. Shah, H. Yokoyama, and N. Kakimoto. High-precision forecasting model of solar irradiance based on grid point value data analysis for an efficient photovoltaic system. *IEEE Transactions on Sustainable Energy*, 6(2):474–481, 2015.

C. Sigauke. Forecasting medium-term electricity demand in a south african electric power supply system. *Journal of Energy in Southern Africa*, 28 (4):54–67, 2017.

C. Sigauke and A. Bere. 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, 2017.

C. Sigauke and D. Chikobvu. Daily peak electricity load forecasting in south africa using a multivariate non-parametric regression approach. *ORiON*, 26(2), 2010.

C. Sigauke and D. Chikobvu. Short-term peak electricity demand in south africa. *African Journal of Business Management*, 6(32):9243–9249, 2012.

C. Sigauke, A. Verster, and D. Chikobvu. Extreme daily increases in peak electricity demand: Tail-quantile estimation. *Energy Policy*, 53:90–96, 2013.

C. Sigauke, Murendeni M. Nemukula, and D. Maposa. Probabilistic hourly load forecasting using additive quantile regression models. *Energies*, 11(9): 2208, 2018.

A.K. Singh, D.E. Allen, and P.J. Robert. Extreme market risk and extreme value theory. *Mathematics and computers in simulation*, 94:310–328, 2013.

G. Singh, D.S. Chauhan, A. Chandel, D. Parashar, and G. Sharma. Factor affecting elements and short term load forecasting based on multiple linear regression method. *International Journal of Engineering Research & Technology (IJERT)*, 3(12), 2014.

S. Sobri, S. Koohi-Kamali, and N.A. Rahim. Solar photovoltaic generation forecasting methods: A review. *Energy conversion and management*, 156: 459–497, 2018.

S. Sun, S. Wang, G. Zhang, and J. Zheng. A decomposition-clustering-ensemble learning approach for solar radiation forecasting. *Solar Energy*, 163:189–199, 2018.

M. Taillardat, A. Fougères, P. Naveau, and R. De Fondeville. Extreme events evaluation using crps distributions. *arXiv preprint arXiv:1905.04022*, 2019.

A. Tilloy, B.D. Malamud, H. Winter, and A. Joly-Laugel. Evaluating the efficacy of bivariate extreme modelling approaches for multi-hazard scenarios. *Natural Hazards and Earth System Sciences*, 20(8):2091–2117, 2020.

C. Voyant, G. Notton, S. Kalogirou, M. Nivet, C. Paoli, F. Motte, and A. Fouilloy. Machine learning methods for solar radiation forecasting: A review. *Renewable Energy*, 105:569–582, 2017.

C. Wan, J. Zhao, Y. Song, Z. Xu, J. Lin, and Z. Hu. Photovoltaic and solar power forecasting for smart grid energy management. *CSEE Journal of Power and Energy Systems*, 1(4):38–46, 2015.

R. Wengenmayr and T. Bührke. *Renewable energy: sustainable energy concepts for the future*. John Wiley & Sons, 2011.

H. Winter. *Extreme value modelling of heatwaves*. PhD thesis, Lancaster University, 2016.

S.N. Wood. Generalized additive models: An introduction with r, 2017.

B. Wright. A review of unit commitment. *ELENE4511, May*, 28, 2013.

T. Xie, G. Zhang, H. Liu, F. Liu, and P. Du. A hybrid forecasting method for solar output power based on variational mode decomposition, deep belief networks and auto-regressive moving average. *Applied Sciences*, 8 (10):1901, 2018.

W. Yamaka, R. Gupta, S. Thongkairat, and P. Maneejuk. Structural and predictive analyses with a mixed copula-based vector autoregression model. *Journal of Forecasting*, 2021.

Z.R. Yet and N. Masseran. Modeling dependence of solar radiation and sky clearness index using a bivariate copula. *Meteorology and Atmospheric Physics*, 133(5):1495–1504, 2021.

B.D. Youngman. Generalized additive models for exceedances of high thresholds with an application to return level estimation for us wind gusts. *Journal of the American Statistical Association*, 114(528):1865–1879, 2019.

B.D. Youngman. Evgam: An r package for generalized additive extreme value models. *arXiv preprint arXiv:2003.04067*, 2020.

M. Zeman. Integrating electricity from solar energy in electricity power system. In *2014 International Conference and Exposition on Electrical and Power Engineering (EPE)*, pages 034–037. IEEE, 2014.

Z. Zhong, C. Yang, W. Cao, and C. Yan. Short-term photovoltaic power generation forecasting based on multivariable grey theory model with parameter optimization. *Mathematical Problems in Engineering*, 2017, 2017.

M. Zieher, M. Lange, and U. Focken. Variable renewable energy forecasting–integration into electricity grids and markets–a best practice guide. In *Proc. Deutsche Gesellschaft Internationale Zusammenarbeit (GIZ)*, pages 1–8, 2015.

Appendix A1: Publications

Research articles (title pages and abstracts only) from this thesis are given in the next three pages.

Mathematics and Statistics 10(1): 116-133, 2022
DOI: 10.13189/ms.2022.100109 http://www.hrpub.org

Estimation of Extreme Quantiles of Global Horizontal Irradiance: A Comparative Analysis Using an Extremal Mixture Model and a Generalised Additive Extreme Value Model

Thakhani Ravele[1,*], **Caston Sigauke**[1], **Lordwell Jhamba**[2]

[1]Department of Mathematical and Computational Sciences, University of Venda, South Africa
[2]Department of Physics, University of Venda, South Africa

Received September 13, 2021; Revised November 2, 2021; Accepted November 21, 2021

Cite This Paper in the following Citation Styles

***(a)**: [1] Thakhani Ravele, Caston Sigauke, Lordwell Jhamba, "Estimation of Extreme Quantiles of Global Horizontal Irradiance: A Comparative Analysis Using an Extremal Mixture Model and a Generalised Additive Extreme Value Model," Mathematics and Statistics, Vol.10, No.1, pp. 116-133, 2022. DOI: 10.13189/ms.2022.100109*

***(b)**: Thakhani Ravele, Caston Sigauke, Lordwell Jhamba, (2022). Estimation of Extreme Quantiles of Global Horizontal Irradiance: A Comparative Analysis Using an Extremal Mixture Model and a Generalised Additive Extreme Value Model. Mathematics and Statistics, 10(1), 116-133. DOI: 10.13189/ms.2022.100109*

Abstract Solar power poses challenges to the management of grid energy due to its intermittency. To have an optimal integration of solar power on the electricity grid it is important to have accurate forecasts. This study discusses the comparative analysis of semi-parametric extremal mixture (SPEM), generalised additive extreme value (GAEV) or quantile regression via asymmetric Laplace distribution (QR-ALD), additive quantile regression (AQR-1), additive quantile regression with temperature variable (AQR-2) and penalised cubic regression smoothing spline (benchmark) models for probabilistic forecasting of hourly global horizontal irradiance (GHI) at extremely high quantiles (τ = 0.95, 0.97, 0.99, 0.999 and 0.9999). The data used are from the University of Venda radiometric in South Africa and are from the period 1 January 2020 to 31 December 2020. Empirical results from the study showed that the AQR-2 is the best fitting model and gives the most accurate prediction of quantiles at τ = 0.95, 0.97, 0.99 and 0.999, while at 0.9999-quantile the GAEV model has the most accurate predictions. Based on these results it is recommended that the AQR-2 and GAEV models be used for predicting extremely high quantiles of hourly GHI in South Africa. The predictions from this study are valuable to power utility decision-makers and system operators when making high-risk decisions and regulatory frameworks that require high-security levels. This is the first application to conduct a comparative analysis of the proposed models using South African solar irradiance data, to the best of our knowledge.

Keywords Additive Quantile Regression, Global Horizontal Irradiance Forecasting, Penalised Cubic Regression Smoothing Spline, Quantile Regression via Asymmetric Laplace Distribution, Semi-parametric Extremal Mixture

1 Introduction

1.1 Background

The generation of power from clean energy sources makes an important contribution to sustainable development. Energy policy-makers are establishing more penetration of clean energy onto the grid, such as solar and wind power [1]. Variability of accurate

STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **10**, September 2022, pp 1235–1249.
Published online in International Academic Press (www.IAPress.org)

Economic dispatch of electrical power in South Africa: An application to the Northern Cape province

Thakhani Ravele [1,*], Caston Sigauke [1], Lordwell Jhamba [2]

[1] *Department of Mathematical and Computational Sciences, University of Venda, South Africa*
[2] *Department of Physics, University of Venda, South Africa*

Abstract Power utility companies rely on forecasting for the operation of electricity demand. This presents an application of linear quantile regression, non-linear quantile regression, and additive quantile regression models for forecasting extreme electricity demand at peak hours such as 18:00, 19:00, 20:00 and 21:00 using Northern Cape data period 01 January 2000 to 31 March 2014. The variables were selected using the least absolute shrinkage and selection operator. Additive quantile regression models were found to be the best-fitting models for hours 18:00 and 19:00, whereas linear quantile regression models were found to be the best-fitting models for hours 20:00 and 21:00. Out of sample forecasts for seven days (01 to 07 April 2014) were used to solve the unit commitment problem using mixed-integer programming. The unit commitment problem results showed that using all the generating units such as hydroelectric, wind power, concentrated solar power and solar photovoltaic is less costly. This study's main contribution is the development of models for forecasting hourly extreme peak electricity demand. These results could be useful to system operators in the energy sector who have to maintain the minimum cost by scheduling and dispatching electricity during peak hours when the grid is constrained due to peak load demand.

Keywords Additive quantile regression, Lasso, Linear quantile regression, Mixed-integer linear programming, Non-linear quantile regression, Unit commitment.

DOI: 10.19139/soic-2310-5070-1057

1. Introduction

1.1. Background

Economic dispatch is essential in power system operation and is defined as the power planning operation with minimum operating costs [1]. The purpose of economic dispatch is to provide optimal power generation at a minimum cost of operation. It also provides the important aspects of power system operation such as meeting load demand at minimum cost by scheduling the committed generating units, reducing the emissions, maintaining the system stability, and security restriction [2]. Electricity load forecasting is important for economic dispatch because it provides future electricity production and consumption [3], which helps electricity utility to maintain the balance of demand and supply [4]. Electricity load forecasting is also important for production planning and trading on the electricity markets. It has different implementations such as energy acquiring and production, load switching, contract rating and infrastructure evolution. Load forecasting helps electric utility management in planning the distribution of electricity [5, 6, 7]. Electricity load forecasting faces rising challenges due to innovative technologies such as smart grids, electric cars, and renewable energy production. The purpose of

*Correspondence to: Thakhani Ravele (Email: ravelethakhani@gmail.com). Department of Mathematical and Computational Sciences, Faculty of Science, Engineering and Agriculture, University of Venda, Private Bag X5050, Thohoyandou, 0950, South Africa.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

energies

MDPI

Article

Extremal Dependence Modelling of Global Horizontal Irradiance with Temperature and Humidity: An Application Using South African Data

Caston Sigauke [1,†], Thakhani Ravele [1,*,†] and Lordwell Jhamba [2]

[1] Department of Mathematical and Computational Sciences, University of Venda, Private Bag X5050, Thohoyandou 0950, South Africa
[2] Department of Physics, University of Venda, Private Bag X5050, Thohoyandou 0950, South Africa
* Correspondence: ravelethakhani@gmail.com; Tel.: +27-76-664-6979
† These authors contributed equally to this work.

Abstract: The interaction between global horizontal irradiance (GHI) and temperature helps determine the maximum amount of solar power generated. As temperature increases, GHI increases up to the point that it increases at a decreasing rate and then decreases. Therefore, system operators need to know the maximum possible solar power which can be generated. Using the multivariate adaptive regression splines, extreme value theory and copula models, the present paper seeks to determine the maximum temperature that will result in the generation of the maximum GHI ceteris paribus. The paper also discusses extremal dependence modelling of GHI with temperature and relative humidity (RH) at one radiometric station using South African data from 16 November 2015 to 16 November 2021. Empirical results show that the marginal increases of GHI converge to 0.12 W/m^2 when temperature converges to 44.26 °C and the marginal increases of GHI converge to −0.1 W/m^2 when RH converges to 103.26%. Conditioning on GHI, the study found that temperature and RH variables have a negative extremal dependence on large values of GHI. Due to the nonlinearity and different structure of the dependence on GHI against temperature and RH, unlike previous literature, we use three Archimedean copula functions: Clayton, Frank and Gumbel, to model the dependence structure. The modelling approach discussed in this paper could be useful to system operators in power utilities who must optimally integrate highly intermittent renewable energies on the grid.

Keywords: archimedean copulas; bivariate threshold excess model; conditional multivariate extreme value; extreme value mixture model; global horizontal irradiance; laplace margins

Citation: Sigauke, C.; Ravele, T.; Jhamba, L. Extremal Dependence Modelling of Global Horizontal Irradiance with Temperature and Humidity: An Application Using South African Data. *Energies* **2022**, *15*, 5965. https://doi.org/10.3390/en15165965

Academic Editor: Ignacio Mauleón

Received: 13 July 2022
Accepted: 14 August 2022
Published: 17 August 2022

1. Introduction

1.1. Context

Sustainable energy supply can be achieved through renewable energy such as solar energy, which is still underused. Solar energy is environmentally friendly since there are no greenhouse emissions when electricity is generated using this resource. This reduces the effects of climate change. The challenge in using renewable energies such as solar is that they are highly intermittent [1]. Solar energy is projected to make up a significant part of the world's energy supply in the future. Accurate forecasting is needed to model intermittent solar energy spatially and temporally. Integrating renewable energy sources into the grid is likely to be one of the most significant challenges of the future world's energy supply [2]. Solar power forecasting is essential for power system operation and adds ancillary services during times of high penetration of renewable energy to avoid variability and uncertainty problems [1]. Due to the intermittence, instability and randomness, it is not easy to produce accurate, reliable forecasts of solar irradiance that can be used as a guide for operating modern smart grids for electricity generation and distribution. Global horizontal irradiance (GHI) hitting the ground surface is the same as electricity generated by photovoltaic (PV)