University of Venda

PROBABILISTIC SOLAR POWER FORECASTING USING PARTIALLY LINEAR ADDITIVE QUANTILE REGRESSION MODELS: AN APPLICATION TO SOUTH AFRICAN DATA

By

Mpfumali Phathutshedzo
11631287

Dissertation submitted for Masters of Science degree in Statistics

in the

Department of Statistics
School of Mathematical and Natural Sciences
University of Venda
Thohoyandou, Limpopo
South Africa

Supervisor: **Dr C Sigauke**

Co-Supervisor: **Dr A Bere**

Co-Supervisor: **Mrs T.S Mulaudzi**

April 2019

Declaration

I Mpfumali Phathutshedzo (11631287) declare that the dissertation for the Master of Science degree in Statistics entitled "Probabilistic solar power forecasting using partially linear additive quantlie regression models: an application to South Africa data", at the University of Venda hereby submitted by me has not been previously submitted for a degree at this or any other university. I further declare that it is my own work in design and in execution and that all referenced material contained therein have been dully acknowledged.

Signature:.....

Date:.....

Dedication

This work is dedicated to my mother Mpfumali MM, my sister Mpfumali R and my best friend Mudzudzanyi T.

Acknowledgment

I would like to take this opportunity to express my deep gratitude to Dr C Sigauke, my research supervisor, for his patient guidance, enthusiastic encouragement and useful critiques of this research work. Without him this work would never have come into existence. My grateful thanks are also extended to Dr A Bere and Mrs S Mulaudzi, my co-supervisors for their useful and constructive recommendations on this research. I would also like to thank Mr Ravele T and Ms Nevhungoni T for their support and also those who directly and indirectly helped me during this research. Finally, I wish to thank my parents for their support and encouragement throughout my study.

Abstract

This study discusses an application of partially linear additive quantile regression models in predicting medium-term global solar irradiance using data from Tellerie radiometric station in South Africa for the period August 2009 to April 2010. Variables are selected using a least absolute shrinkage and selection operator (Lasso) via hierarchical interactions and the parameters of the developed models are estimated using the Barrodale and Roberts's algorithm. The best models are selected based on the Akaike information criterion (AIC), Bayesian information criterion (BIC), adjusted R squared ($AdjR^2$) and generalised cross validation (GCV). The accuracy of the forecasts is evaluated using mean absolute error (MAE) and root mean square errors (RMSE). To improve the accuracy of forecasts, a convex forecast combination algorithm where the average loss suffered by the models is based on the pinball loss function is used. A second forecast combination method which is quantile regression averaging (QRA) is also used. The best set of forecasts is selected based on the prediction interval coverage probability (PICP), prediction interval normalised average width (PINAW) and prediction interval normalised average deviation (PINAD). The results show that QRA is the best model since it produces robust prediction intervals than other models. The percentage improvement is calculated and the results demonstrate that QRA model over GAM with interactions yields a small improvement whereas QRA over a convex forecast combination model yields a higher percentage improvement. A major contribution of this dissertation is the inclusion of a non-linear trend variable and the extension of forecast combination models to include the QRA.

Keywords: *Lasso via hierarchical interaction, Partially linear additive models, Probabilistic solar power forecasting, Forecast combination.*

Table of contents

List of Figures

List of Tables

Abbreviations

$AdjR^2$	Adjusted R Squared.
AIC	Akaike Information Criteria.
AICc	Akaike Information Criteria corrected.
AnEn	Analog Ensemble.
ANN	Artificial Neural Networks.
BIC	Bayesian Information Criteria.
BR	Barrodale and Roberts.
CV	Cross Validation.
DHI	Direct Horizontal Irradiance.
DiffHI	Diffuse Horizontal Irradiance.
ECMWF	European Center for Medium-range Weather Forecasts.
ELM	Extreme Learning Machine.
ESC	Electricity Supply Commission.
ETS	Exponential Smoothing.
GAM	Generalised Additive Model.
GAMs	Generalised Additive Models.
GAMMs	Generalised Additive Mixed Models.
GCV	Generalised Cross Validation.
GEFCom	Global Energy Forecasting Competition.
GHI	Global Horizontal Irradiation.
GLM	Generalised Linear Model.
GSET	Group for Solar Energy Thermodynamics.
KSI	Kolmogorov-Smirnon test Integral.
LASSO	Least Absolute Shrinkage and Selection Operator.
LOOCV	Leave One Out Cross Validation.
LTSF	Long-Term Solar power Forecasts.
MAE	Mean Absolute Error.
MAPD	Mean Absolute Percentage Deviation.

MAPE	Mean Absolute Percentage Error.
MaxAE	Maximum Absolute Error.
MSE	Mean Square Error.
MBE	Mean Bias Error.
MTSF	Medium-Term Solar power Forecasts.
NRMSE	Normalised Root Mean Square Error.
NWP	Numerical Weather Prediction.
OLSR	Ordinary Least Squares Regression.
PeEn	Persistence Ensemble.
PI	Prediction Interval.
PICP	Prediction Interval Coverage Probability.
PINAD	Prediction Interval Normalised Average Deviation.
PINAW	Prediction Interval Normalised Average Width.
PIW	Prediction Interval Width.
PLAMs	Partially Linear Additive Models.
PLAQR	Partially linear additive quantile regression.
PSF	Probabilistic solar power forecasting.
PV	Photovoltaic.
QR	Quantile regression.
QRA	Quantile Regression Averaging.
R^2	Coefficient of Determination.
RMSE	Root Mean Squared Error.
RSS	Residual Sum of Squares.
SA	South Africa.
SAURAN	South African Universities' Radiometric Network.
STSF	Short-Term Solar power Forecasts.
SVMs	Support Vector Machines.
WRF	Weather Research and Forecasting.
W/m^2	Watts Per Square Meter.

List of Notation and Special Symbols

$F_{Y\|X}$	Conditional distribution function of X given X.
$F^{-1}_{Y\|X}$	Inverse function of the conditional distribution function of Y given X.
$Q_{Y\|X}$	Conditional quantile function.
$Q_{0.5}(Y\|X)$	Median quantile regression function.
s_j	Smooth functions.
$\hat{Q}_y\|X, Z(\tau)$	Average loss function.
ρ_τ	Pinball loss function.
β_0	Intercept.
β_j	Coefficients.
$\mathbf{B}$	Model matrix of basis functions.
$\mathbf{X}_t$	Covariates matrix.
θ	Parameter vector.
$\ell(\theta)$	Maximised log-likelihood.
$\varepsilon_{t,\tau}$	Random error term.

Chapter 1

Introduction

South Africa (SA) currently benefits from clean electricity produced by renewable energy technologies such as solar and wind. Such technologies have their own advantages and can be implemented depending on local conditions. For instance, the coastal areas are windy, implying suitability for wind energy. In urban areas solar is a better option than wind because the turbines are noisy. Also compared to wind farms, solar does not require much space (Warner, 2014).

All renewable energies are advantageous in contrast with burning coal for energy since fossil fuels produce climate-changing carbon emissions and hence they are not sustainable energy sources. Fossil fuels are also getting more expensive due to carbon emission levels whereas renewable technologies such as solar panels are getting cheaper over the long term because the sun's energy is free. That surely appeals to every profit making business, more especially as the cost of electricity in SA continues to rise (Warner, 2014).

In this dissertation the focus is on solar power modelling and forecasting. For the rest of this dissertation the term solar forecasting is used to refer to solar power forecasting. (PSF) will be used as the abbreviation for probabilistic solar power forecasting. Probabilistic forecasting takes probability distributions over future events into account. They are preferred over point forecasting because they take into account the uncertainties in the predicted values which helps in taking effective decisions. Therefore it is believed that probabilistic forecasting can become a power trend in solar power forecasting (Mohammed et al., 2015).

Across the world, solar energy technologies are developing robustly. Solar power applications are useful in the construction of solar plants. Such facilities have huge potential for utilizing the energy in sunny regions (Adnan et al., 2015). Medium and long term solar power forecasts are typically derived from numerical weather prediction (NWP) models but statistical and machine learning techniques are increasingly being used together with the NWP models to produce more accurate forecasts since measurements (weather stations) are available at a few distinct spatial points only. The data used in this study from Eskom. Eskom is SA's electricity public utility established by the government of the Union of SA as the Electricity Supply Commission (ESC) in 1923.

1.1 Statement of the problem

Solar power forecasting is an activity which has increased in both interest and importance as a result of the increasing penetration of solar power into

the global energy mix (Huang and Perry, 2016).

Just like electric load forecasts, solar power forecasts are also assessed periodically, ranging from minutes, hours, days, weeks and months up to a year. These periods are classified as follows, short-term solar power forecasts (STSF), medium-term solar power forecasts (MTSF) and long-term solar power forecasts (LTSF). Short term forecasts are from few hours up to a week and medium term forecasts are from weeks up to a year whereas long term forecasts are for more than a year (Fan and Hyndman, 2012). This dissertation will focus on medium-term forecasting of global horizontal irradiance (GHI) data from Tellerie radiometric station in South Africa. GHI is the total solar radiation on a horizontal surface and is the sum of the direct horizontal irradiance (DHI) plus the diffuse horizontal irradiance (DiffHI) and the Albedo. Forecasting of GHI is the first and most essential step in most PV power prediction system. GHI forecasting approaches may be categorized according to the input data used which also determine the forecast horizon (Antonanzas et al., 2016).

Partially linear additive quantile regression (PLAQR) models are used to forecast the hourly solar power irradiance in Northern Cape in SA. The PLAQR models are a combination of generalised additive models and QR. This dissertation aims to predict solar power generation for some of the solar plants located in SA.

In SA many studies regarding solar power/energy have been done (i.e. Dekker et al., 2015; Le Roux, 2016; Kruger and Sekele, 2013), but none of

them used QR. Juban et al. (2016) proposed a generic framework based on a multiple quantile regression (MQR) for probabilistic energy forecasting whereas Bacher et al. (2009) suggested a two-stage method basing the clear sky model on statistical smoothing techniques and quantile regression.

1.2 Purpose of the study

1.2.1 Aim

The main aim of this study is to develop probabilistic medium-term solar power forecasting models for predicting solar power generation at Tellerie radiometric station in SA.

1.2.2 Objectives

The objectives of the study are to:

- develop partially linear additive quantile regression models for probabilistic solar power forecasting,
- estimate the parameters of the developed models using the Barrodale and Roberts algorithm for ℓ_1-regression,
- evaluate the performance of the models,
- forecast the hourly solar irradiance,
- evaluate the accuracy of the forecasts.

1.2.3 Scope of the dissertation

QR will be used to predict hourly global solar irradiance using the selected SA's solar data. Hourly solar power data will be obtained from Eskom. PLAQR models will be used for a probabilistic solar power forecasting. The parameters of the PLAQR models will be estimated using the Barrodale and Roberts algorithm for ℓ_1-regression. The Barrodale and Roberts algorithm for ℓ_1-regression is known to be efficient for problems up to several thousands of observations. It can be used to compute the full quantile regression process and also uses an efficient scheme for computing confidence intervals. The best model will be selected using information criteria: Akaike information criterion (AIC), AIC corrected(AICc), Bayesian information criterion (BIC), Cross validation (CV) and adjusted R squared ($AdjR^2$). The accuracy of the forecasts will be evaluated using mean absolute percentage error (MAPE), mean absolute error (MAE) and root mean square errors (RMSE).

1.2.4 Significance of the study

When large solar farms are integrated into the power grid, solar power forecasting becomes important for system operators who make decisions about the power grid operations as well as for the electric market operators. This study will help the system operators to take optimal decisions in the scheduling and dispatching of electricity to consumers.

1.2.5 Structure of the dissertation

The rest of the dissertation is organized as follows: Chapter 2 contains literature review on PSF and the summary of studies using methods similar to

the proposed one. Chapter 3 provides general theory of the QR and PLAQR statistical methods that will be used in this study. Chapter 4 provides the data analysis and the conclusion are given in Chapter 5.

Chapter 2

Literature Review

2.1 Introduction

Solar power is an energy source with a promising future. It has the ability to meet all energy demands around the world. This chapter provides an overview of solar power (SP) and the summaries of some studies used in the proposed methodology to forecast solar irradiance in different countries.

2.2 An overview of solar power forecasting

On the solar panels the magnitude of solar irradiation is directly proportional to solar power output. The integration of high penetration of solar energy generation requires accurate forecasting. Cloud movement, cloud formation, and cloud dissipation are the main causes of solar irradiance variations. The effect of clouds on solar irradiance drop depends on the cloud's height and thickness. Thick clouds absorb radiation, but they can also scatter radiation from their sides so that some extra radiation gets to the radiometers resulting

in enhancement. Thin clouds scatter the radiation so that much of it becomes diffuse rather than direct and this affects results in enhancement.

Researchers have developed different methods for solar power forecasting, such as statistical approaches using historical data, the use of numerical weather prediction (NWP) models, tracking cloud movements from satellite images, and tracking cloud movements from direct ground observations using sky cameras (Zhang et al., 2015).

2.2.1 Probabilistic solar power forecasting

Different models for solar forecasting have been developed and used to predict solar irradiance with the aim of finding the model that will produce the best probabilistic forecasts of solar power.

Nonparametric additive model

Huang and Perry (2016) use a nonparametric approach to develop a probabilistic solar power forecasting model. The authors use the proposed approach to calculate the prediction intervals using a k-nearest neighbour's regression model. They use the historical data from the ECMWF numerical weather prediction model comprising the actual solar power and twelve independent variables. The developed model proves to be successful in producing accurate forecasts of solar power outputs together with realistic prediction intervals.

Grantham et al. (2016) propose a new method for constructing a full predictive density of solar irradiance based on a nonparametric bootstrap. The bootstrap method is a general tool to assess statistical accuracy. The authors demonstrate a probabilistic forecasting of solar irradiance on an hourly basis. They express the hourly GHI for Mildura as follows:

$$I_t = F_t + A_t + Z_t, \tag{2.2.1}$$

where F_t denote a cyclic component, A_t an autoregressive component, and Z_t an error term. The results show that the new approach is computationally inexpensive and easily tractable and it also delivers sharp and calibrated ensembles of one-hour ahead forecasts. The authors further suggest that for future improvements the nonparametric bootstrap approach should be extended to a case of autoregressive conditionally heteroscedastic models for tracking non-linear dynamics in solar irradiance using sieve and other types of approximations as well as to a more general case of joint multivariate wind and solar probabilistic forecasting.

Another study which applies a nonparametric model is that of Faranak et al. (2016). The authors propose a nonparametric density forecasting method based on an extreme learning machine (ELM) as a regression tool, trained with an appropriate criterion. The authors apply the proposed nonparametric method on two PV power datasets with one-minute resolution and highly fluctuating patterns. The results show that the proposed method is able to efficiently provide reliable and sharp predictive densities for the very short-term such as (10-minute and one-hour lead times) forecasts. The authors further suggest that for future improvements the nonparametric approach

could be extended to account for NWP as input when looking at higher lead times.

Quantile regression model

Quantile regression (QR) is an approach that has been employed in several studies (Bremnes, 2004; Nielsen et al., 2006; among others) in order to create nonparametric probabilistic forecasts. Quantile regression provides the conditional quantile of an output value given a certain input and it is obtained through the minimization of a non-differentiable loss-function (Fahrmeir et al., 2009). QR models are used in forecasting by different energy sectors such as solar power forecasting, wind power forecasting, etc.

Bacher et al. (2009) propose an online power forecasting from the PV systems approach. The authors use the PV systems approach to predict hourly values of solar power for horizons of up to 36 hours. They use 15-minute observations of solar power data from 21 PV systems located on rooftops in a small village in Denmark. The authors suggest a two-stage method where they first use a clear sky model to obtain a statistical normalization of solar power and then the adaptive linear time series models to calculate the forecasts of the normalised solar power. They based a clear sky model on a statistical smoothing techniques and QR. The results indicate that for forecasts of up to 2 hours ahead the available observations of solar power are the most important input, while for longer horizons NWPs are the most important input.

Alessandrini et al. (2015) generate probabilistic solar power forecasts (PSF) using the application of an analog ensemble (AnEn) method. The authors based the method on an historical set of deterministic NWP model forecasts and observations of the solar power. They compare the AnEn performance for PSF with the QR technique and a persistence ensemble (PeEn) over three solar farms in Italy. One of the advantages of these methods (the AnEn, QR and PeEn) is that they do not require an irradiance-to-power conversion function. The data used is hourly power data from three solar farms located in Italy. The results show that AnEn gives the best level of statistical consistency compared to QR and PeEn. The authors suggest that the tests they carried out could be expanded to other locations spanning a wider range of climatological conditions in future.

Juban et al. (2016) propose a generic framework for probabilistic energy forecasting based on multiple quantile regression and also discuss the application of the method to several tracks in the 2014 Global Energy Forecasting Competition (GEFCom2014). The authors use the approach to predict a full distribution over possible future energy outcomes. As a result the method is ranked on the top five for the three of the four tasks (solar, wind and price) and is ranked the seventh in the load forecasting task.

This dissertation will use PLAQR models which are a combination of quantile regression and generalised additive models. Variable selection will done using the least absolute shrinkage and selection operator (Lasso) via hierarchical interactions. Lasso will be used as a model selection that addresses problems of multicollinearity. In the following section we briefly discuss lit-

erature in which Lasso has been used in regression based models.

Lasso for variable selection

Variable selection in regression models is important as it helps in reducing the dimension of the problem and at the same time avoids over-fitting. Lasso which was developed by Tibshirani (1996) is used in this dissertation for variable selection. Lasso uses ℓ_1 penalty. It is one of the shrinkage methods that include elastic net and ridge regression. In the literature there are a few cases where the shrinkage method Lasso has been used in solar power forecasting models.

Yang et al. (2015) propose the use of Lasso to perform a sub-5-minute irradiance forecasting. This is used to automatically shrink and select the most appropriate lagged time series for regression. It is found that the proposed model outperforms univariate time series models and ordinary ridge regression significantly and that it has a good performance when the training data are few and predictors are many. In this dissertation Lasso via hierarchical interactions (Bien et al., 2013; Lim and Hastie, 2015) will be used to select the appropriate variables. Lasso via hierarchical interactions is based on a strong hierarchy constraint (Bien et al., 2013; Lim and Hastie, 2015). The strong hierarchy constraint is one in which if variables are included in the interactions, then the cross-effects will also be included in the main effects (Bien et al., 2013; Lim and Hastie, 2015).

Quantile regression for partially linear additive models and generalised additive models

Hu et al. (2015) propose a Bayesian quantile regression method for partially linear additive models. The authors use the simulation studies to illustrate the algorithm. They further illustrate the proposed approach using two real data sets. The approach developed by Hu et al. (2015) for partially linear additive models (PLAMs) is given in the following equation:

$$\mathrm{y}_i = \mu_\tau + \sum_{j=1}^{p} f_{\tau,j}(x_{ij}) + \varepsilon_{\tau,i}, \qquad i = 1, ..., n, \tag{2.2.2}$$

where (y_i, x_i) are independent and identically distributed pairs, y_i is the response, $x_i = (x_{i1}, ..., x_{ip})^T$ is the p dimensional predictor vector, μ_τ is the intercept, ε_i is random errors with τth quantile equal to 0 and $f_{\tau,j}$ is a univariate component function; it might either be linear, non-linear or zero. The simulation results show that the methods perform well even though the errors are not generated based on their assumed generating mechanism.

Hastie and Tibshirani (1986) introduce a class of generalised additive models (GAMs) that replace the linear form $\sum \beta_j X_j$ by the sum of smooth functions $\sum s_j(x_j)$ where $s_j(.)$ denotes the unspecified functions which are estimated using scatter plots smoothers. They illustrate the technique using the binary response and survival data and the method proves to be useful in uncovering non-linear covariates effects.

Another study by Shadish et al. (2014) apply the GAMs and generalised additive mixed models (GAMMs) to single-case design data. They find that

these models excel at detecting whether the trend exists or not. They also show how GAMMs can be used to estimate the autoregressive model and random effects and finally discuss the limitations of the mixed models in comparison with GAMs.

This dissertation will apply both the partially linear additive quantile regression and generalised additive models to forecast solar irradiance addressing autocorrelation in the residuals. To reduce the autocorrelation, seasonal autoregressive integrated moving average (SARIMA) models will be fitted to the residuals.

SARIMA models

Bouzerdoum et al. (2013) introduces a hybrid model for short-term power forecasting of a grid-connected PV plant. The model combines two methods: the seasonal auto-regressive integrated moving average method (SARIMA) and the support vector machines method (SVMs). An experimental database of the power produced by a small-scale 20 kWp GCPV plant is used to develop and verify the effectiveness of the proposed model in short-term forecasting. Hourly forecasts of the power produced by the plant are carried out for a few days giving accurate forecasts. Empirical results show that the developed hybrid model performs better than both the SARIMA and the SVM model.

Kardakos et al. (2013) address two practical methods for electricity generation forecasting of grid-connected PV plants. The first model is based on SARIMA time-series analysis and is further improved by incorporating

short-term solar radiation forecasts derived from NWP models. The second model adopts artificial neural networks (ANN) with multiple inputs. Day-ahead and rolling intra-day forecast updates are implemented to evaluate the forecasting errors. All models are compared in terms of the normalised (with respect to the PV installed capacity) Root Mean Square Error (NRMSE). Test results from the application of all models in different PV plants of the Greek power system show that, in general, those models that make use of available external solar radiation data, namely the modified SARIMA model and the ANN models are preferable, since they lead to considerably improved day-ahead forecasts and lower NRMSEs as compared to the persistence model or the pure SARIMA model.

Aloysius et al. (2015) propose weather research and forecasting (WRF) model to forecast day-ahead solar irradiance using data from a tropical environment in Singapore. The WRF model is benchmarked using persistence and two seasonal time series models, namely, the exponential smoothing (ETS) and seasonal autoregressive integrated moving average (SARIMA) models. The WRF model outperforms the SARIMA model and achieves accuracies comparable with persistence and ETS models. The persistence, ETS, and WRF models have relative RMSE of about 5557%. Furthermore, they find that by combining the forecasting outputs of WRF and ETS models, errors can be reduced to 49%.

Forecast combination

It has been established that combining forecasts produces more accurate forecasts (Bates and Granger, 1969; Devaine et al., 2012; Gaillard, 2015).

Mohamed (2017) uses the ensemble learning tool and random forest to combine the individual models to obtain hour-ahead combined point forecasts and then generate the ensemble based probabilistic solar power forecasts. The author concluded with an overall evaluation that the ensemble based-probabilistic forecasts are more accurate than the analog-ensemble and persistence probabilistic forecasts and that the random forecasting procedure is a powerful ensemble learning method for combining the different forecasts and quantifying the probability distribution of these combined forecasts.

Performance evaluation

Coimbra and Marguez (2013) propose metric for evaluating the quality of solar forecasting models, defined as the ratio of solar uncertainty to solar variability. They use the proposed metric to evaluate two new forecasting models and compare their performance with a persistence model. They also compare the results of the proposed forecasting metric to the conventional forecasting performance such as the RMSE, and find that the proposed metric is a good indicator of the extent to which a forecast model effectively predicts the stochastic variability of solar irradiance.

Another study by Zhang et al. (2013b) develop a suite of generally applicable, value based metrics for solar forecasting for a comprehensive set

of scenarios that can evaluate the economic and reliability impact of improved solar forecasting. The proposed metrics are divided into five categories namely (i) statistical metrics, including Pearsons correlation coefficient, root mean square error (RMSE), maximum absolute error (MaxAE), mean absolute error (MAE), mean absolute percentage error (MAPE), mean bias error (MBE), KolmogorovSmirnov test Integral (KSI), skewness, and kurtosis; (ii) variability estimation metrics, including different time and geographic scales, and distributions of forecast errors; (iii) uncertainty quantification and propagation metrics, including standard deviation and information entropy of forecast errors; (iv) ramping characterization metrics, including swinging door algorithm signal compression and heat maps and (v) economic and reliability metrics, including non-spinning reserves service represented by 95^{th} percentiles of forecast errors. To evaluate the performance of the proposed metrics they use the actual and forecast solar power data from western wind and solar integration study and find that the developed metrics can successfully evaluate the quality of a solar forecast.

The RMSE and MAE will be used in this dissertation to evaluate the performance of the developed models. RMSE penalises large errors and avoids taking absolute values which is known to be undesirable in many mathematical calculations. The RMSE is on the same scale as the data. On the other hand, MAE is easy to understand, explain, interpret and compute.

Forecast error distribution

Zhang et al. (2013a) investigate the correlation between wind and solar power forecast errors. The forecast and actual data is obtained from the western wind and solar integration study. They analyse both the day-ahead and 4-hour-ahead forecast errors for the western interconnection of the United States. They use the kernel density estimation method to estimate a joint distribution of both wind and solar power forecast errors. They also evaluate the Pearson's correlation coefficient between wind and solar forecast error. Their results show that wind and solar power forecast errors are weakly correlated. This dissertation will use the kernel density and summary statistics (mean, median, skewness and kurtosis) to estimate the distribution of solar irradiance forecast errors.

2.3 Conclusion

To the best of our knowledge the use of partially linear additive quantile regression models and generalised additive models on solar probabilistic solar power forecasting in South Africa has not been done to date. Also there are few previous studies where quantile regression models were applied to solar power forecasting but none were based on quantile regression only.

Chapter 3

Methodology

3.1 Introduction

This chapter discusses the general theory of the proposed methods, i.e. generalised additive models, quantile regression models and the partially linear additive quantile regression models. A generalised additive model (GAM) which was first introduced by Hastie and Tibshirani (1990) follows from the additive models. It has the relation to a generalised linear model (GLM) but additionally the linear predictors include a sum of smooth functions of the independent variables and a consistent parametric component of the linear prediction. Henceforth GAM incorporates both parametric and nonparametric model components and the model can indicate the dependence of the response on the independent variable in a flexible way, not relying upon the assumption that all relations can be modelled as linear. The general formula for GAM is as follows (Hastie and Tibshirani, 1990):

$$g(\mu_t) = \mathbf{X_t}\boldsymbol{\theta} + \sum_{j=1}^{k} s_j(x_{tj}) + \varepsilon_t, \quad t = 1, ..., n, \tag{3.1.1}$$

where $\mu_t = E(Y_t)$ and $Y_t \sim$ some exponential family distribution, Y_t is a response variable, $\mathbf{X_t}$ represents the covariates matrix, $\boldsymbol{\theta}$ is the parameter vector and s_j are smooth nonparametric functions of the covariates x_{tj}. Generally there are three types of smoothers used for GAM and are defined as follows (Larsen, 2015):

- Local regression (Loess): Loess has a place in the class of nearest neighbourhood-based smoothers. With a specific end goal to acknowledge Loess, we need to comprehend the most simplistic member of this family which is the running mean smoother. Running mean smoothers are symmetric, moving averages. Loess creates a smoother curve than the running mean by fitting a weighted regression inside each nearest-neighbour window, where the weights depend on a kernel that suppresses data points far away from the objective data point.

- Smoothing splines: They adopt a totally different strategy in deriving smooth curves. As opposed to utilizing a nearest-neighbour moving window, the smooth function was estimated by minimizing the penalized sum of squares. The function that minimizes the penalized sum of squares is a natural cubic spline with knots at every data point, which is otherwise called a smoothing spline. However, for predictive modelling, smoothing splines have a major disadvantage: it is not practical to have knots at every data point when dealing with large models.

- Regression splines (B-splines, P-splines, thin plate splines): They offer a more practical alternative to smoothing splines. The main advantage of regression splines is that they can be expressed as a linear com-

bination of a finite set of basis functions that do not depend on the dependent variable Y which is practical for prediction and estimation. Regression splines of order q can be written as follows:

$$\mathrm{s}(x) = \sum_{i=1}^{k} \mathrm{B}_{i,q}(x)\beta_i, \tag{3.1.2}$$

where $\mathrm{B}_{1,q}(x), ..., \mathrm{B}_{k,q}(x)$ are basis functions, q denotes the basis dimension, B is the model matrix of basis functions, and $\beta = [\beta_1, ..., \beta_p]^T$ are the coefficients. The number of basis functions depends on the number of inner knots a set of ordered, distinct values of x_j as well as the order of the spline. Specifically, if we let m to denote the number of inner knots, then the number of basis functions is given by $k = p + 1 + m$. In this dissertation the regression splines (B-splines to model the non-linear relationships) are used. B-spline functions are piecewise polynomial functions and they provide more flexibility than power series (Wood, 2017). They are also more stable in a sense that each B-spline is non-zero over a limited range of knots (Wood, 2017).

3.2 Quantile Regression

Quantile regression (QR) is defined as a statistical technique that allows the estimation of the conditional quantile function. It is also used to obtain statistical inference about the quantiles in the same way as classical regression methods based on minimizing sum of squares of residuals that facilitate estimation of conditional mean functions (Hardle et al., 2000) are used. The QR model was first introduced by Koenker and Bassett (1978) as an exten-

sion of median regression (the ℓ_1 estimator). The ℓ_1 estimator is analogous to ordinary least squares regression (OLSR) which is mostly used to establish a relationship between a response (dependent) variable and one or more independent variables (Davino et al., 2014).

QR has several advantages: It does not have limitations like the conditional mean as it provides location, scale, and shape of the entire conditional distribution of the response variable given a set of predictors (Davino et al., 2014).

Suppose Y is a random variable (i.e. hourly solar power irradiance) and X is a set of explanatory (independent) variables (i.e. surface solar radiation, total column liquid water of cloud, surface pressure, etc.), then the conditional quantile Q_τ of order $\tau \in (0, 1)$ of Y given X is defined as the inverse of $F_{Y|X}$, if and only if $F_{Y|X}$ is the conditional distribution function of Y given X (Davino et al., 2014):

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

The idea of quantile estimation arises from the observation that minimizing the expected absolute error yields the median (i.e., $q_{o.5}(Y|X)$. It can be shown that the conditional quantile $Q_\tau(Y|X)$ is the solution of the following minimization problem:

$$Q_\tau(Y|X) \in \arg\min_g \mathbb{E}[\rho_\tau(Y - g(X))|X], \tag{3.2.2}$$

where, ρ_τ is a pinball loss ($\rho_\tau(u) = u(\tau - I(u < 0))$ defined for all $u \in \Re$ (Gaillard et al., 2016). Equation (3.2.2) can be written as follows:

$$Q_\tau(Y|X) = \arg\min_g \sum_{y_i > g(X_i)} \tau(y_i - g(X_i)) + \sum_{y_i \leq g(X_i)} (1-\tau)(y_i - g(X_i)). \quad (3.2.3)$$

3.3 Partially linear additive quantile regression

3.3.1 Additive models

A simpler strategy for dimension reduction in multivariate nonparametric models involves additive models of the form (Koenker, 2005):

$$Q_\tau(Y|X) = \beta_0 + \sum_{j=1}^{p} g_j(X^T\beta_j) + \varepsilon_t, \quad (3.3.1)$$

where Y is the response variable (solar irradiance), X are the independent variables, β_0 is the intercept, β_j are parameters and g_j is an unknown function. A partially linear model (PLAM) with a nonparametric component and an additive linear parametric component are regarded as the most important special examples of the additive models (Koenker, 2005).

3.3.2 Partially linear additive models

PLAMs are regarded as a special case of GAMs. The PLAMs are more flexible than stringent linear models since they have both linear and nonlinear additive components and they are also more parsimonious than the general nonparametric regression models (Hoshino, 2014). The general formula for

PLAMs is as follows:

$$\mathrm{y}_t = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \sum_{j=1}^{q} s_j(zj) + \varepsilon_t, \qquad t = 1, ..., n. \tag{3.3.2}$$

Given a quantile level $\tau \in (0, 1)$, equation (3.3.2) can be extended to a PLAQR model and has the form:

$$\mathrm{y}_{t,\tau} = \beta_{0,\tau} + \sum_{j=1}^{p} \beta_{j,\tau} x_j + \sum_{j=1}^{q} s_{j,\tau}(z_j) + \varepsilon_{t,\tau}, \qquad t = 1, ..., n, \quad \tau\epsilon(0, 1), \tag{3.3.3}$$

where $y_{t,\tau}$ is the response variable, x_j are linear variables, $\beta_{0,\tau}$ is the intercept, $s_{j,\tau}$ are smooth functions, z_j are non-linear variables, $\beta_{j,\tau}$ are parameters and $\varepsilon_{t,\tau}$ are random errors, the total number of parameters will be $k = p + q + 1$. Equation (3.3.3) can be written in a vector form:

$$Y = s_\tau(X) + Z^T \beta_\tau + \varepsilon. \tag{3.3.4}$$

Minimizing the following function

$$Q(\beta_\tau, s_\tau(.)) = \sum_{j=1}^{n} \rho_\tau(Y_j - Z_j^T \beta - s(X)) + \sum_{j=1}^{n} \lambda_j \int (s''(t))^2 dt \tag{3.3.5}$$

gives the estimators of equation (3.3.4). The non-linear relationships in the PLAQR models are modelled through the use of B-splines. B-splines are known to be very stable and flexible basis for large scale spline interpolation (Wood, 2017).

3.4 Parameter estimation

This study will use an adaptation of the Barrodale and Roberts (BR) algorithm for ℓ_1-regression($\tau = 0.5$) to estimate the PLAQR models' parameters

for all quantile levels $\tau \in (0,1)$. It exploits the characteristics of the ℓ_1 approximation in order to solve the problem in a more efficient manner than the general simplex approach, although it is a simplex based method. The BR's algorithm is based on the linear programming technique. Linear programming is a mathematical programming technique used to optimize an objective function subject to a set of constraints. The simplex approach to solving the general ℓ_1-regression problem is given as follows:

$$\min_{\beta \epsilon \Re^p} \sum_{j=1}^{n} \rho_\tau (Y_j - Z_j^T \beta - s(X)). \tag{3.4.1}$$

Expression (3.4.1) can be reformulated as a linear programming problem (equation 3.4.2) by introducing $2n$ artificial or (slack) variables $u_j, v_j : 1, ..., n$ representing the positive and negative parts of the vector of residuals. The BR's algorithm is then used to solve the following problem (Portnoy and Koenker, 1997):

$$\min\{\mathbf{1}^T \mathbf{u} + \mathbf{1}^T \mathbf{v} | y = \mathbf{Z}\beta + \mathbf{u} - \mathbf{v}, (\mathbf{u}, \mathbf{v}) \epsilon \Re_+^{2n}\}, \tag{3.4.2}$$

where $\mathbf{1}$ is a vector of ones and $\mathbf{Z}$ denotes the n by p regression matrix.

3.5 Variable and model selection

3.5.1 Variable selection

Variable selection is necessary because most models don't deal well with a large number of irrelevant variables. With such large number of variables, there might exist problems among explanatory variables, in particular there could be a problem with multicollinearity. Multicollinearity is the condition

where two or more predictors variables in a statistical model are linearly related (Dormann et al., 2013). Also with a large number of variables there is often a desire to select a smaller subset that not only fits as well as the full set of variables but also contains the more important variables. This helps to reduce both the dimension of the problem and over-fitting.

In this dissertation variables will be selected using shrinkage methods. A shrinkage method is a general procedure that is used in improving a least squares estimator which comprises in reducing the variance by adding constraints on the value of the coefficients (Bontempi and Ben, 2011). There are three types of shrinkage methods, namely: ridge regression, Lasso and elastic net. In this study the focus is on the use of Lasso in variable selection. Lasso is a useful variable selection method, especially when dealing with highly correlated predictors and it allows for automatic variable selection.

Lasso via hierarchical interactions

Elastic net is a shrinkage method that includes ridge regression and Lasso shrinkage (here ridge regression and Lasso shrinkage are special cases of elastic net). In order to achieve better prediction in the face of multicollinearity, Hoerl and Kennard (1970) proposed ridge regression which minimizes residual sum of squares (RSS) subject to:

$$\sum_{i=1}^{p} |\beta_j|^2 \leq t \qquad (\ell_2 \text{ Norm}), \tag{3.5.1}$$

For the problem of multicollinearity ridge regression improves the prediction performance, but it cannot produce a model with only the relevant variables,

unlike Lasso. Lasso is defined as a regression method that is used in variable selection. It uses the ℓ_1 penalty loss function:

$$\hat{\beta}_{Lasso}(\lambda) = \arg\min ||\vec{y} - \mathbf{X}\hat{\beta}||_2^2 + \lambda||\hat{\beta}||_1. \tag{3.5.2}$$

Interactions are an important part of the regression model in such a way that adding them to a regression model can greatly expand the understanding of the relationship among the variables. Since there are $\binom{p}{k}$ interactions of order k, fitting regression models with interactions is quite challenging, more especially when one has even moderate number (p) of measured variables. This study focuses on pairwise ($k = 2$) interaction models (Bien et al., 2013).

Pairwise interaction models

Consider a regression model with pairwise interactions between the predictors $x_1, ..., x_P$ (i.e. total column liquid water of cloud, surface pressure, etc.) for an outcome variable Y (i.e. hourly global solar irradiance). In particular the pairwise interaction model will have the form (Bien et al., 2013):

$$y_{t,\tau} = \beta_{0,\tau} + \sum_j \beta_{j,\tau} X_j + \frac{1}{2}\sum_{j \neq k} \Theta_{jk,\tau} X_j X_k + \varepsilon_{t,\tau}, \tag{3.5.3}$$

where $\beta_{0,\tau}$ is the intercept, $\beta_{j,\tau}$ are the parameters, $\Theta_{jk,\tau}$, the second term after the intercept β_0 is referred to as the main effects and the third as the interaction terms. The factor of half before the interaction summation is a consequence of notational decision to deal with a symmetric matrix Θ of interactions rather than a vector of length $p(p-1)/2$ (Bien et al., 2013). When fitting model (3.5.3) the interactions are allowed into the model only if at least one of the corresponding main effects is in the model. There are

two types of hierarchical restrictions, namely: strong hierarchy (In strong hierarchy an interaction model is said to obey a strong hierarchy if and only if both its main effects are present) and weak hierarchy (A weak hierarchy is said to be obeyed only if one of its main effects is present) (Bien et al., 2013). Therefore the PLAQR model with hierarchical pairwise interactions can be written as:

$$y_{t,\tau} = \beta_0 + \sum_{j=1}^{p} \beta_{j,\tau} x_j + \sum_{j=1}^{q} s_{j,\tau}(z_j) + \sum_{j \neq i} \frac{1}{2}(\gamma_{ji,\tau} x_j x_i) + \sum_{j \neq i} \frac{1}{2}(\sigma_{ji,\tau} x_j z_i) + \sum_{j \neq r} \frac{1}{2}(\phi_{jr,\tau} z_j z_r) + \varepsilon_{t,\tau} \quad (3.5.4)$$

and the error terms may be autocorrelated, i.e. they follow a SARIMA(p,d,q)×(P,D,Q)[s] model given in equation 3.5.5.

$$\phi(\beta)\Phi(\beta^s)\nabla^d\nabla^D\varepsilon_{t,\tau} = \theta(\beta)\Theta(\beta^s)a_t. \quad (3.5.5)$$

Now let

$$m(x,z) = \sum_{j=1}^{p} \beta_{j,\tau} x_j + \sum_{j=1}^{q} s_{j,\tau}(z_j) + \sum_{j \neq i} \frac{1}{2}(\gamma_{ji,\tau} x_j x_i) + \sum_{j \neq i} \frac{1}{2}(\sigma_{ji,\tau} x_j z_i) + \sum_{j \neq r} \frac{1}{2}(\phi_{jr,\tau} z_j z_r) + \varepsilon_{t,\tau}, \quad (3.5.6)$$

then

$$\mathrm{y}_{t,\tau} = \beta_0 + m(x,z) + \varepsilon_{t,\tau} \quad (3.5.7)$$

implies

$$\varepsilon_{t,\tau} = \mathrm{y}_{t,\tau} - \beta_0 - m(x,z). \quad (3.5.8)$$

Now substituting in equation (3.5.5) results in equation 3.5.9

$$\phi(\beta)\Phi(\beta^s)\nabla^d\nabla^D[\mathrm{y}_{t,\tau} - \beta_0 - m(x,z)] = \theta(\beta)\Theta(\beta^s)a_t. \quad (3.5.9)$$

3.5.2 Model selection

The best model is selected based on the Akaike information criterion (AIC), corrected AIC (AICc), Bayesian information criterion (BIC), cross validation (CV) and adjusted R squared ($AdjR^2$). The AIC was first developed by Akaike (1973).

AIC, AIC Corrected and BIC

AIC score aims to find the approximate model whereas the BIC score aims to find the true model. When comparing AIC values for multiple models, smaller values of the criterion are better. The AIC is computed as follows:

$$\text{AIC} = -2k - 2\log \ell(\Theta). \tag{3.5.10}$$

When comparing BIC values for multiple models, the best model is the one that provides the minimum BIC. The BIC is calculated as follows:

$$\text{BIC} = k\log(n) - 2\log \ell(\Theta), \tag{3.5.11}$$

where $\log \ell(\Theta)$ is the value of the maximized log-likelihood, k is the number of parameters, Θ is the set (vector) of model parameters and n the number of observations (Fabozzi et al., 2014).

AICc is defined as the AIC with a correction for finite sample sizes. AICc is the AIC with greater penalty for extra parameters. The formula for AICc is as follows:

$$\text{AICc} = \text{AIC} + \frac{2k(k+1)}{n-k-1}, \tag{3.5.12}$$

where n denotes the sample size and k the number of parameters (Fabozzi et al., 2014).

Cross validation

Cross validation (CV) is used to partition a sample of data into two subsets, a training (calibration) subset for model fitting and a test (validation) subset for model evaluation. This approach is efficient on a large sample size. There are several types of cross-validation namely: the k-fold cross-validation, the 2-fold cross validation also known as the holdout method, repeated random sub-sampling (the sample data is first shuffled and then used as cross validation and training data) and the leave-one-out cross validation (LOOCV) (Hastie et al., 2001). The leave-one-out is extended to generalised cross validation (GCV). This study makes use of leave-one-out and GCV to select the best model. LOOCV is a very general method which can be used with any kind of predictive modelling. A model with a minimum cross-validation value is preferred.

The leave-one-out cross-validation

In leave-one-out, only a single record from a sample data is used for cross validation. The method is fit on the $n-1$ training observation and using the value x_i^T to predict $\hat{y}_i$, where x_i^T denotes the ith row vector of x_i^T and $\hat{y}_i$ denotes that observation is left in parameter estimation is made for the excluded observations. The MSE_i is then computed on the excluded observations $(x_i^T, \hat{y}_i)$, that is $\text{MSE}_i = (y_i - \hat{y}_i)^2$. The MSE1 provides an unbiased estimate for the test error. This is repeated n times resulting in n squared errors, $\text{MSE}_1, ..., \text{MSE}n$. The leave-one-out CV estimate for the test MSE is

computed as follows (James et al., 2013):

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}\mathrm{MSE}_i. \tag{3.5.13}$$

Generalised cross validation

Generalised cross validation (GCV) was first developed by Wahba (1990). It is a generalisation of the leave-one-out cross validation approach. The strategy is to delete data points, one at a time, then fit a smoother to the remaining data points followed by fitting a smoother against the whole dataset (Larsen, 2015). It measures the goodness of fit which considers the residual error and the model complexity. GCV criterion is given by equation (3.5.14) (Craven and Wahba, 1979):

$$\mathrm{GCV(M)} = \frac{\frac{1}{n}\sum_{t=1}^{n}[y_t - \hat{f}_{\mathrm{M}}(\mathbf{X}_t)]^2}{[1 - \frac{G(\mathrm{M})}{n}]^2}, \tag{3.5.14}$$

where n denotes the sample size, and G(M) is the cost-penalty measure of a model containing M basis functions. The numerator determines the lack of fit on the M basis function model $\hat{f}_M(\mathbf{X}_t)$ and the denominator represents the penalty for the model complexity G(M). The best model is one with the lowest GCV criterion value.

Adjusted R-squared

A modified version of R-squared that has been adjusted for the number of predictors in the model is called the $AdjR^2$. Whenever there is a new term that improves the model more than it was expected, the $AdjR^2$ increases and decreases when a predictor improves the model by less than expected chance.

The $AdjR^2$ is always lower than the R-squared. Using the $AdjR^2$, the best model will be the one with the highest value of $AdjR^2$. The $AdjR^2$ is defined as follows:

$$Adj\mathrm{R}^2 = 1 - \left[\frac{(1-R^2)(n-1)}{n-k-1}\right], \tag{3.5.15}$$

where n is the number of points in the data sample and k is the number of independent regressors (number of variables in the model) (Hyndman and Athanasopoulos, 2014).

3.6 Residual analysis

After fitting a model it is important to determine whether all the necessary assumptions of the residuals are valid before performing inference because if there are any violations, the subsequent inferential procedures may be invalid resulting in faulty conclusions. One of the tests will be to check whether or not the residuals are autocorrelated. Residuals are defined as the difference between the actual and the fitted values and are expressed as follows:

$$\varepsilon_{t,\tau} = y_t - \hat{y}_t, \qquad t = 1, ..., n, \tag{3.6.1}$$

where y_t are the actual values of a response variable and $\hat{y}_t$ are the estimated values of a response variable (Hyndman and Athanasopoulos, 2014).

3.6.1 Autocorrelations

When fitting a model the autocorrelation of the residuals also has to be considered and accounted for. When the original data is highly autocorrelated, the residuals have a high risk of being autocorrelated. Thus, autocorrelation

has to be dealt with, however this can be done by adding more parameters in the fit. Reducing autocorrelation by adding more parameters is quite hard. Therefore a SARIMA model can be added to the model. A formal test for autocorrelation procedure can be summarized as follows (Hyndman and Athanasopoulos, 2014; Marasinghe, 2014):

- Estimate the parameters.
- Check if the response variables and predictors are stationary. If not apply differencing until all variable become stationary.
- Extract residuals and use the Ljung Box-test to test for autocorrelation.
- If still autocorrelated, identify an appropriate SARIMA(p,d,q)×(P,D,Q)[s] model by repeating the process until the desired results are achieved.

3.6.2 SARIMA model

The general SARIMA model can be represented as follows:

$$\phi(\mathbf{L})\Phi(\mathbf{L}^s)\nabla^d\nabla_s^D z_t = \theta(\mathbf{L})\Theta(\mathbf{L}^s)a_t \sim N(0,\sigma^2) \tag{3.6.2}$$

where z_t represents solar irradiance, $a_t \sim N(0,\sigma^2)$ is error term, s is the seasonal length, $\mathbf{L}$ is a backshift operator (or lag operator) ($\mathbf{L}zt = z_{t-1}$). $\phi(\mathbf{L}) = (1-\phi_1\mathbf{L}-...-\phi_P\mathbf{L}^p)$ is the non-seasonal autoregressive (AR) operator, $\Phi(\mathbf{L}^s) = (1 - \Phi_1\mathbf{L}^s - ... - \Phi_P\mathbf{L}^{Ps})$ is the seasonal AR operator, $\theta(\mathbf{L}) = (1 - \theta_1(\mathbf{L}) - ... - \theta_q\mathbf{L}^q)$ is the non-seasonal moving average (MA) operator, $\Theta(\mathbf{L}^s) = (1 - \Theta_1(\mathbf{L}^s) - ... - \Theta_Q\mathbf{L}^{Qs})$ is the seasonal MA operator. ∇^d and ∇^D are the non-seasonal and seasonal difference operators of order d and D respectively, where $\nabla^d = (1-\mathbf{L})^d$ and $\nabla^D = (1-\mathbf{L}^s)^D$.

3.7 Model evaluation

The accuracy of the forecasts model will be evaluated using mean absolute percentage error (MAPE), mean absolute error (MAE) and root mean square error (RMSE).

3.7.1 Mean absolute error and root mean square error

MAE measures the average of the absolute differences between prediction and actual observation where all individual differences have equal weight while RMSE is the square root of the average of squared differences between prediction and actual observation. They both measure the average magnitude of the error. The two measures are statistically presented as follows:

$$\text{MAE} = \frac{1}{m}\sum_{t=1}^{m}|y_t - \hat{y}_t|, \tag{3.7.1}$$

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{t=1}^{m}(y_t - \hat{y}_t)^2}, \tag{3.7.2}$$

where m is the number of observations in the test data set, $\hat{y}_t$ is the estimated values of a response variable and y_t is the actual values of the t^{th} observation.

3.7.2 Mean absolute percentage error

MAPE (also known as mean absolute percentage deviation (MAPD)) measures the percentage of average absolute error that has occurred. It expresses the accuracy as percentages and is defined as:

$$\text{MAPE} = \frac{100}{m}\sum_{t=1}^{m}\left|\frac{y_t - \hat{y}_t}{\hat{y}_t}\right|\%, \tag{3.7.3}$$

where y_t is the actual value and $\hat{y}_t$ is the forecast value (Adhikari and Agrawal, 2013).

3.7.3 Forecast combination

Combining forecasts was first introduced by Bates and Granger (1969). It is particularly useful when one is uncertain about a situation, uncertain about which method is most accurate and when one wants to avoid large errors. It improves the forecast accuracy to the extent that the component contains useful and independent information.

The loss functions play important roles in forecast combination in two ways; the first one is that it may serve as a key ingredient in combination formulas and the second one is that they are used to define performance evaluation criteria. In the first direction the use of loss function is found in many popular combination schemes such as regression based combination and many adaptive forecast combination methods.

Let $y_{t,\tau}$ be hourly global solar irradiance and let there be K methods used to predict the next m observations of $y_{t,\tau}$, i.e. m is the total number of forecasts. The combined forecasts are then given by equation (3.7.4) (Gailard and Goude, 2015; Sigauke, 2017):

$$z_{t,\tau} = \sum_{k=1}^{K} w_{t,k,\tau}\hat{y}_{t,k} + \varepsilon_{t,\tau}, \quad \tau \in (0,1), \quad t = 1, ..., m, \tag{3.7.4}$$

where $w_{t,k,\tau}$ is the weight assigned to the forecast $\hat{y}_{t,k}$.

3.7.4 Quantile regression averaging

Quantile regression averaging (QRA) was first introduced by Nowotarski and Weron (2015) as a new approach to compute the prediction intervals. It can be viewed as an extension of combining forecasts. It involves applying QR to the forecasts of a small number of individual forecasting models. This approach treats the individual forecasts as the independent variables and the corresponding observation as the dependent variable (global solar irradiance). Let $y_{t,\tau}$ be hourly global solar irradiance as discussed in Section 3.7.3, then combined forecasts will be given by

$$\hat{y}_{t,\tau}^{\text{QRA}} = \beta_{0,\tau} + \sum_{k=1}^{K} \beta_{k,\tau}\hat{y}_{t,k} + \varepsilon_{t,\tau}, \quad \tau \in (0,1), \quad t = 1, ..., m, \tag{3.7.5}$$

where $\hat{y}_{t,k}$ represents forecasts from method k, $\hat{y}_{t,\tau}^{\text{QRA}}$ is the combined forecasts and $\varepsilon_{t,\tau}$ is the error term. We seek to minimise

$$\arg\min_{\beta} \sum_{t=1}^{m} \rho_\tau\big(\hat{y}_t^{\text{QRA}} - \beta_0 - \sum_{k=1}^{K} \beta_k \hat{y}_{t,k}\big). \tag{3.7.6}$$

In matrix form, we have (Maciejowska et al., 2017):

$$\arg\min_{\beta \in IR} \sum_{t=1}^{m} \rho_\tau\big(\hat{y}_t^{\text{QRA}} - x_t^T\beta\big), \tag{3.7.7}$$

which reduces to

$$\arg\min_{\beta \in IR^\rho} \sum_{t:\hat{y}_t^{\text{QRA}} > x_t^T\beta} \tau\big(\hat{y}_t^{QRA} - x_t^T\beta\big) + \sum_{t:\hat{y}_t^{\text{QRA}} \leq x_t^T\beta} (1-\tau)(\hat{y}_t^{\text{QRA}} - x_t^T\beta). \tag{3.7.8}$$

The QRA forecasts will be compared with forecasts based on weighted average of the forecasts given in equation (3.7.4).

3.7.5 Pinball loss

The accuracy of the probabilistic forecasts are evaluated using the pinball loss function, which is defined as a piecewise linear function which is used to evaluate the accuracy of quantile forecasts and is given by equation (3.7.7) (Abuella et al., 2015):

$$L(\hat{Q}_{Y|X}) = \begin{cases} \tau(y - \hat{Q}_{Y|X}) & \text{if} \quad Y \geq \hat{Q}_{Y|X} \\ (1-\tau)(\hat{Q}_{Y|X-y}) & \text{if} \quad Y < \hat{Q}_{Y|X} \end{cases}, \tag{3.7.9}$$

where $\hat{Q}_{Y|X}$ is the loss function to the probabilistic forecast and y is the observed value of solar irradiance. To yield more accurate forecasts the average of the loss function $\hat{Q}_{Y|X}$ for all forecasting hours is minimized. The average loss function value is then taken as a score to evaluate the model performance.

3.7.6 Prediction interval widths

Prediction Interval Width (PIW) measures the extension of the interval as the difference of the estimated upper and lower limit values. Generally it is given by the following equation (3.7.8) (Sun et al., 2017):

$$\text{PIW}_i = UL_i - LL_i \qquad i = 1, ..., k, \tag{3.7.10}$$

where UL_i and LL_i is the upper and lower limit. The PI normalised Average Width (PINAW) and PI Coverage Probability (PICP) are used to evaluate the performance of prediction intervals (PIs). The PINAW describes the width of the PIs and the PICP evaluate the reliability of the PIs and are

given by equation (3.7.9) and (3.7.10):

$$\text{PINAW} = \frac{1}{m}\sum_{i=1}^{m}\frac{UL(X_i) - LL(X_i)}{y_{max} - y_{min}}, \tag{3.7.11}$$

$$\text{PICP} = \frac{1}{m}\sum_{i=1}^{m} I_i \qquad I_i = \begin{cases} 1 & \text{if} \quad y_i \epsilon (L_i, U_i) \\ 0 & \text{otherwise} \end{cases}, \tag{3.7.12}$$

where y_{min} and y_{max} represent the minimum and the maximum values of PIW respectively. For the PIs to be valid, the PICP value has to be greater than or equal to the predetermined confidence level. The deviation degree from the actual value to the PIs is described by the PI normalised Average Deviation (PINAD). The PINAD is defined as:

$$\text{PINAD} = \frac{1}{m}\sum_{i=1}^{m}\frac{d_i}{y_{max} - y_{min}} \qquad d_i = \begin{cases} L_i - y_i & \text{if} \quad y_i < L_i \\ 0 & \text{if} \quad L_i \leq y_i \leq U_i \\ y_i - U_i & \text{if} \quad y_i > U_i \end{cases}. \tag{3.7.13}$$

3.7.7 Forecast error distributions

Mean, median, Q_1 (first quartile) and Q_3 (third quartile) are frequently used in the characterization of the forecast error distribution. They provide important information about the distribution. However skewness and kurtosis can provide more information. Skewness measures the probability distribution's asymmetry. Kurtosis describes the magnitude of the distribution's peak. It also measures the thickness of the distribution tails. A distribution with a high kurtosis value is leptokurtic and the one with a low kurtosis value is platykurtic (Hodge et al., 2012).

3.7.8 Percentage improvement

The percentage improvement between the best model with the other models is calculated using equation 3.7.12.

$$\text{Improvement} = \left(1 - \frac{\text{pinball(best model)}}{\text{pinball(other model)}}\right) * 100 \tag{3.7.14}$$

where pinball (best model) is the pinball loss of the best model and pinball (other model) is the pinball loss of the other model.

3.8 Summary

This chapter presented the general theory of the proposed methods. The focus was placed on producing a hybrid model called PLAQR which is a combination of both the GAM and QR models. The parameters of this model as indicated in Section 3.5 will be estimated using the BR algorithm. In order to avoid fitting model with a large number of irrelevant variables a variable selection technique called Lasso via hierarchical pair wise interactions will be used to select suitable variables. Section 3.7 also emphasised that it is important to account for autocorrelation of the residuals when fitting a model. QRA model as indicated in Subsection 3.8.4 will be used to combine forecasts and the results will be compared with a convex combination method. The next chapter presents detailed analysis of the data using the methods discussed in this chapter.

Chapter 4

Data analysis

4.1 Introduction

This chapter presents detailed analysis of the data using the methods discussed in Chapter 3. Generalised additive models (GAMs) and partially linear additive quantile regression (PLAQR) models are compared and discussed. The analysis will be done using the R-software.

4.2 Data

4.2.1 Data source

The data used in this study is hourly averaged global horizontal irradiance (GHI) (average W/m^2) measured by Eskom (www.eskom.co.za), at site Tellerie in Northern Cape, South Africa for the period August 2009 to April 2010 giving a sample size of $n = 5698$ observations. The target variable is the GHI (average W/m^2). The covariates used in this study are (these were the only available covariates on the Eskom website): Relative humid-

ity (RH_Avg), Rain fall (RN), Wind speed (WS_ms_S_WVT), Air temperature (AirTC_Avg), Barometric pressure (BP_mbar_Avg) and Wind direction(WindDir_D1_WVT). In addition to the covariates given above, a factor variable (Hour (Hr)), Hour = 1, Hour = 2,..., Hour = 24; a non-linear trend (noltrend) and lagged GHI (Lag1 and Lag2) variables are added in order to model the trend in the time series data and also to help in reducing the autocorrelation, respectively. A non-linear trend is obtained by fitting a penalised cubic smoothing spline to the response variable (GHI). The penalised cubic smoothing spline function is given by equation (4.2.1):

$$\Pi(t) = \sum_{t=1}^{n}(y_t - f(t))^2 + \lambda \int f''(t)^2 d(t), \tag{4.2.1}$$

where λ is the smoothing parameter which is selected based on the generalised cross validation (GCV) criterion. Lag1 and Lag2 represent the lagged global solar irradiance of the first and second previous hours, therefore taking the lags reduced the sample size from 5700 to 5698.

The wind speed and wind direction are both measured at 9m levels. All the variables provided have an impact on the solar power generated. There are many predictors of solar irradiance that are left out in this study such as surface pressure, total cloud cover, total precipitation, etc. These variables were not available on the Eskom website. The exact location of Tellerie radiometric station is:

- Latitude: $27°22.505'S$
- Longitude: $21°17.799'E$

Figure 4.1 presents a map of South Africa showing the location of Eskom solar measuring station Tellerie located in Northern Cape.

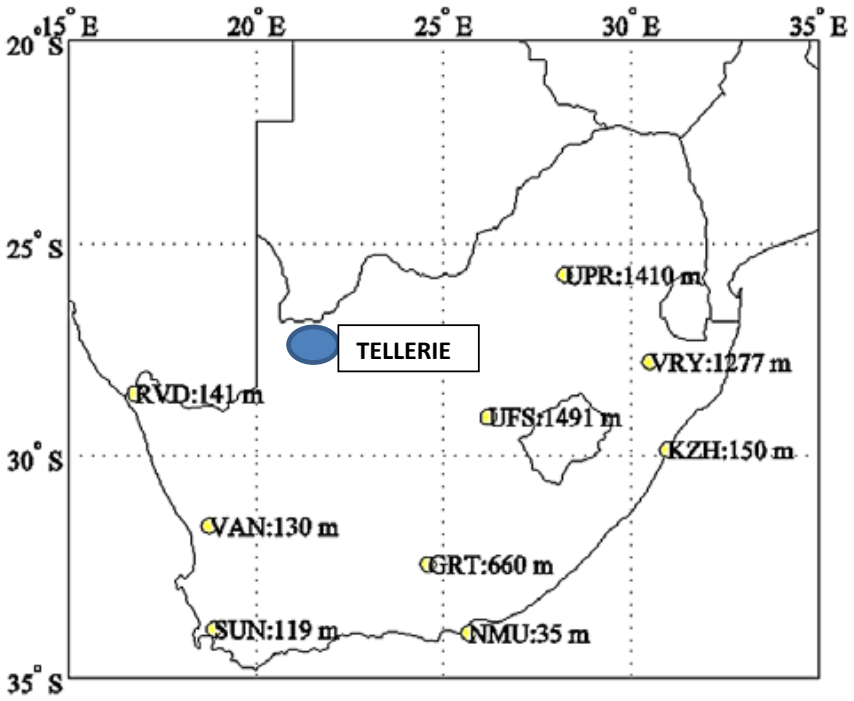

Figure 4.1: Map of South Africa showing the location of the Eskom solar measuring station Tellerie located in Northern Cape (Zhandire, 2017).

The data for the period 28 August 2009 (00:00hrs) to 8 April 2010 (10:00hrs) is used for training ($n_1 = 5362$ observations), while the remaining data i.e. 9 April 2010 (11:00hrs) to 22 April 2010 (09:00hrs) ($n_2 = 336$ observations) are used for testing. The training and testing data sets are used for all the models developed in this study.

4.2.2 Exploratory data analysis

It is considered important to first get the analysis of historical data before setting up the forecasting models. The time series, density, box and normal Q-Q plots for the dependent variable (GHI (average W/m^2)) are given in Figure 4.2. A time series plot for GHI superimposed with the 1^{st} and the 99^{th} quantile forecasts is provided in Appendix A, Figure 1. The data is divided into two subsets: a training set and a testing set to build the base forecasting methods and evaluate the performance. Plot (a) in Figure 4.2 displays global solar irradiance (average W/m^2 in South Africa measured hourly from 2009-2010. There is an upward trend from observation 1-3000, followed by a downward trend from observation 3000-5700 (Figure 4.2 (a)). Plot (c) shows a non-linear relationship between the theoretical and sample quantiles of GSI is seen in plot (c) suggesting that GSI data are not normally distributed and also confirmed by plots (b) and (d).

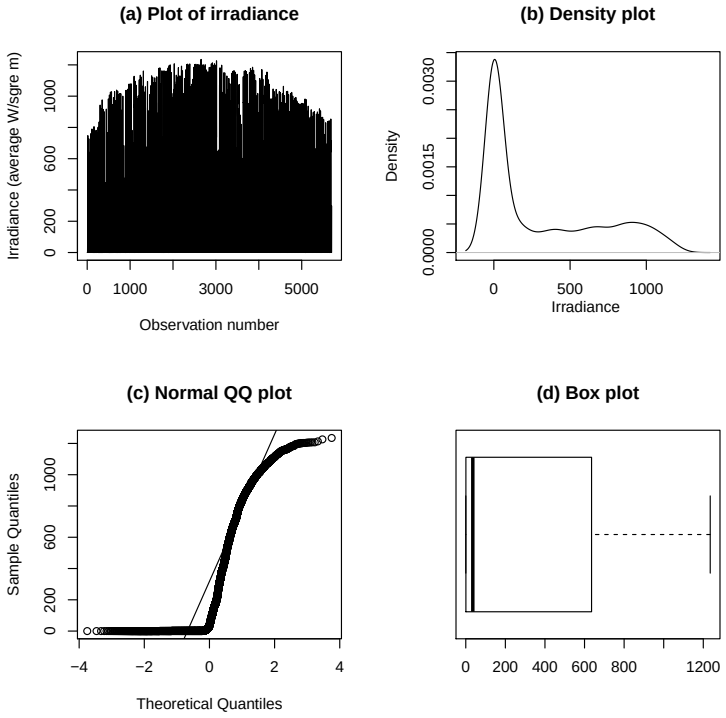

Figure 4.2: Diagnostic plots for global solar irradiance (average W/m^2).

Summary statistics of the global solar irradiance

Table 4.1: Summary statistics of the global solar irradiance (average W/m^2).

Min	1st Qu	Median	Mean	3rd Qu	Max	Skewness	Kurtosis
0.000	0.64	34.84	305.91	635.20	1235.00	0.86	-0.79

Table 4.1 presents summary statistics of global solar irradiance. Minimum global solar irradiance is 0.000 (average W/m^2) and maximum global solar irradiance is 1235.00 (average W/m^2). The skewness coefficient presented in

Table 4.1 suggests that the distribution of global solar irradiance is skewed to the right. The mean and the median are not equal which confirms that the distribution is not normally distributed.

4.3 Generalised additive models

This section discusses two generalised additive models (GAMs), the one with and without interactions. GAM without interaction is the one with all the predictor variables(main effects), and the GAM with interactions is the one which includes all the significant variables selected using Lasso via hierarchical interactions.

4.3.1 GAM without interaction

The GAM without interaction for estimating the parameters $(\beta_0, \beta_1, ..., \beta_7)$ is given by equation (4.3.1).

$$y_t = \beta_0 + \beta_1 s(\text{RN}) + \beta_2 s(\text{WS}) + \beta_3 s(\text{WindDir}) + \beta_4 s(\text{AirTC}) + \beta_5 s(\text{RH}) + \beta_6 s(\text{BP}) + \beta_7 s(\text{HOUR}) + \varepsilon_t \quad (4.3.1)$$

where y_t is the response variable (global solar irradiance (average W/m^2) and ε_t denotes the error term which is assumed to be autocorrelated. After estimating the parameters, the next step is to check how the covariates are related to the dependent variable. Figures 4.3 and 4.4 allow us to evaluate the nature of the relationship between the covariates and dependent variable.

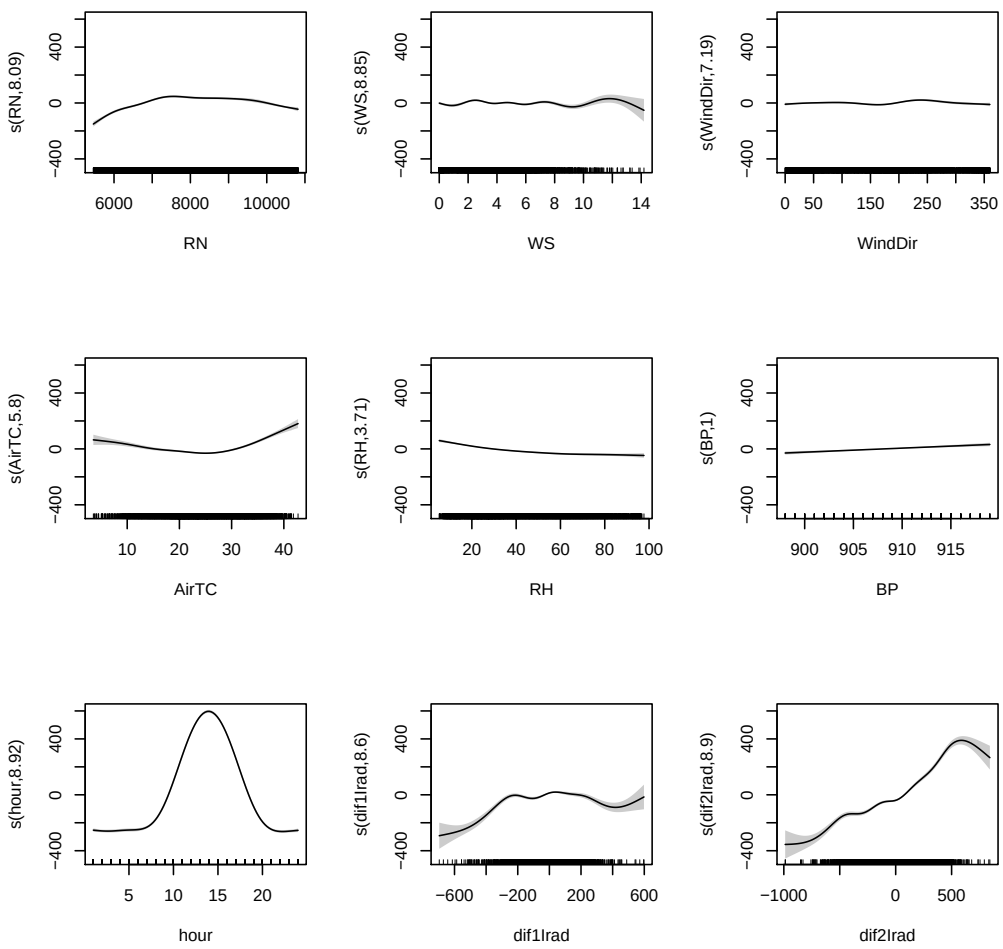

Figure 4.3: For the solar power data, the GAM without interaction model is fitted to the global solar irradiance. Each plot displays the fitted function which is either linear or nonlinear.

The bottom panel of Figure 4.3 indicates that solar irradiance increases over time since there is a peak between 10:00 and 15:00 hours. The relationship between global solar irradiance and BP is linear while with all the other

variables RN,WS, WindDir, AirTC, RH and hour it is nonlinear.

Figure 4.4 displays diagnostic plots for the residuals for the generalised additive model without interactions (from model 4.3.1). Top panel left plot suggests that the residuals are not normally distributed. Right plot shows raw residuals versus the linear predictor and shows no pronounced pattern. Bottom panel right plot shows the sign of over estimation in the upper interval. Left plot, the histogram of residuals suggests that the residuals are not normally distributed.

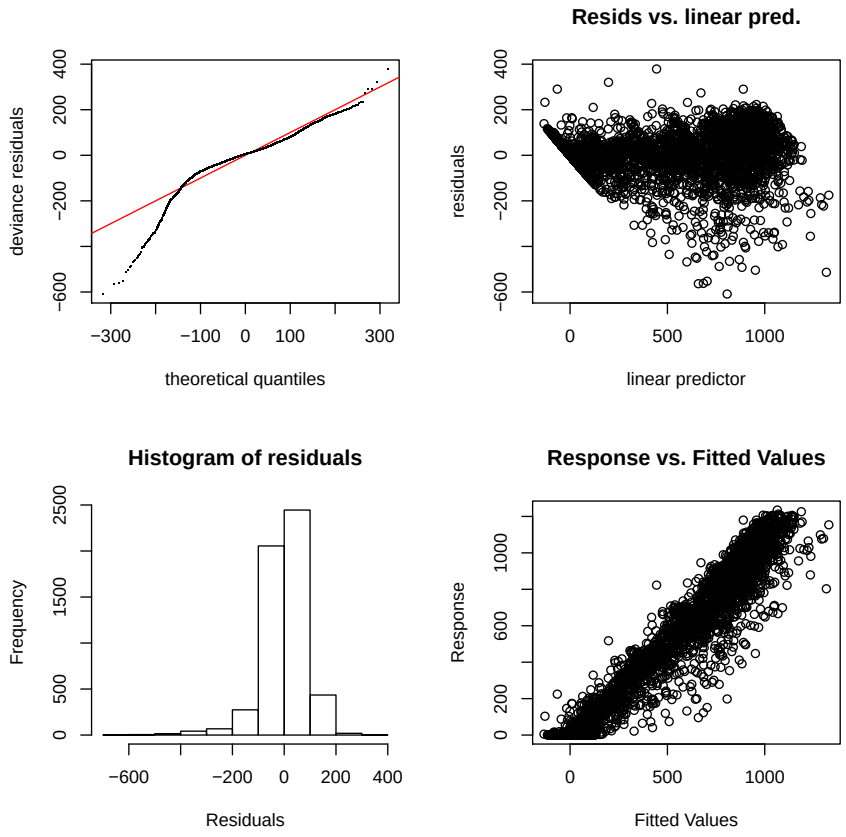

Figure 4.4: Diagnostic plots for the residuals **Top left:** Quantile-Quantile (Q-Q) plot of residuals.

Residual analysis

Since the data used for this analysis are a time series, it is necessary to take account of their autocorrelation structure. In order to check the autocorrelation, the Ljung-Box test is applied to the residuals from the fitted GAM model. The Ljung-Box test statistic for the developed model is 10935, and has a p-value of $< 2.2 \times 10^{-}16$. Since the p-value for the test statistic is

less than 0.05, it can be concluded that the data has significant autocorrelation. To account for the autocorrelation in the residuals, the SARIMA $(2,0,3)(1,0,1)[24]$ model is used:

$$\begin{aligned}\varepsilon_t = &-y_t + \Phi_1 y_{t-24} + \phi_1 y_{t-1} - \phi_1\Phi_1 y_{t-25} + \phi_2 y_{t-2} - \phi_2\Phi_1 y_{t-26} \\ &+ \Theta_1\varepsilon_{t-24} - \theta_1\Theta_1\varepsilon_{t-25} + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} - \theta_1\Theta_1\varepsilon_{t-26} \\ &+ \theta_3\varepsilon_{t-3} - \theta_3\Theta_1\varepsilon_{t-27}. \quad (4.3.2)\end{aligned}$$

Fitting the SARIMA model and applying the Ljung-Box test up to lag 36 results in a p-value of 0.1034 which is greater than 0.05 meaning that there are no autocorrelations left in the residuals at the 5 percent level of significance. Therefore model (4.3.2) is selected to fit the residuals. Figure 4.5 displays the time series of the residuals before and after reducing the autocorrelations.

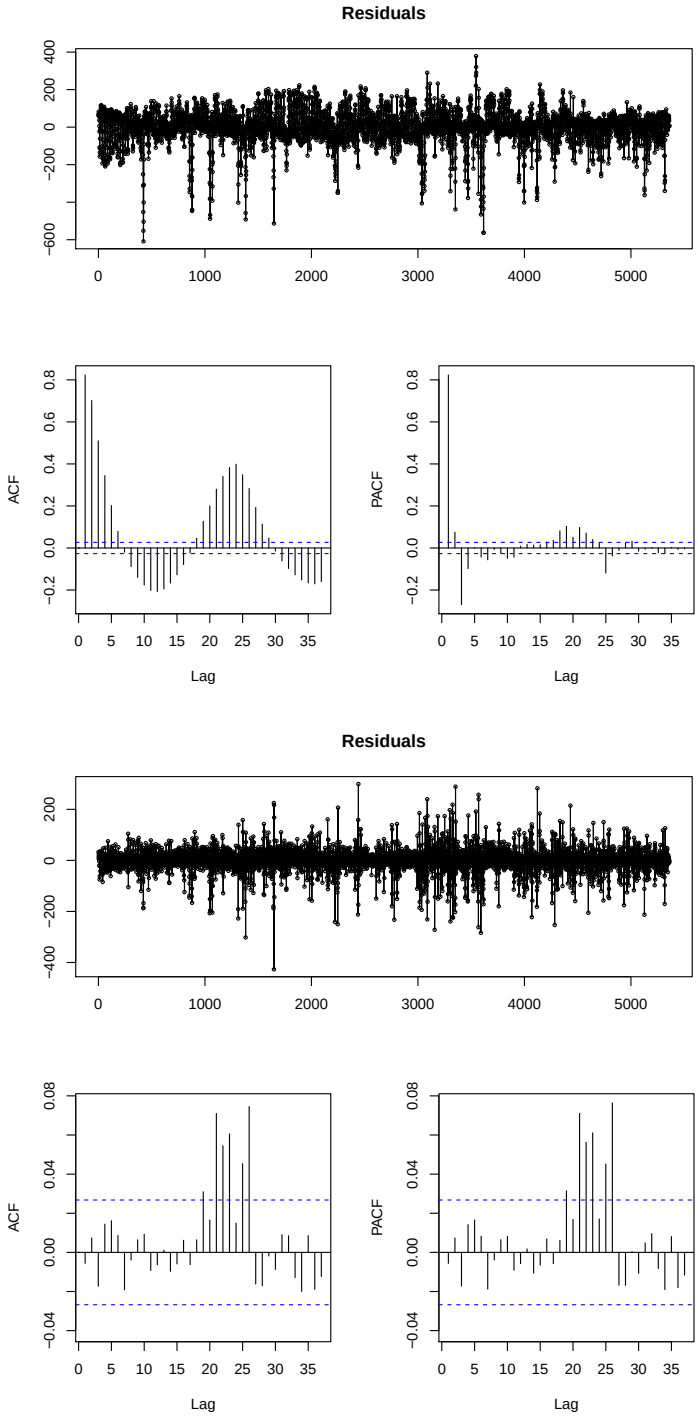

Figure 4.5: **Top**: Time series display for the residuals before reducing the autocorrelation. **Bottom**: Time series display for the residuals after reducing the autocorrelation for model 4.3.2.

4.3.2 Variable selection using Lasso

After accounting for the autocorrelation by a SARIMA model for GAM without interactions, the variable selection using Lasso is done and found that all variables are significant as shown in Table 4.2.

Table 4.2: Variable selected using Lasso.

Variables	Coefficients
(Intercept)	-1.878661×10^{4}
RN	-1.536983×10^{-2}
WS	1.249123×10
WindDir	$2.921835 \times 10^{-}1$
AirTC	3.902018×10
RH	-2.046057×10^{0}
BP	2.011544×10
hour	-6.220940×10^{0}
noltrend	1.425628×10^{-1}
Lag1	-5.464509×10^{-1}
Lag2	9.024672×10^{-1}

4.3.3 Variable selection using Lasso via hierarchical interactions

Variable selection using Lasso via hierarchical interactions is carried out and a summary of the results are presented in Table 4.3.

Table 4.3: Variable selected using Lasso via hierarchical interactions.

	Main effect	**RN**	**WS**	**WindDir**	**AirTC**	**RH**	**BP**	**HOUR**	**noltrend**	**Lag1**	**Lag2**	**Tight?**
RN	-6.29	-4.62	0	0	0	0	0	0	0	0	-1.67	
WS	4.04	0	0	0	0	0	4.04	0	0	0	0	
WindDir	8.87	0	0	-1.14	0	0	4.29	-3.007	0.43	0	0	*
AirTC	169.62	0	0	0	43.81	-15.83	0.61	-106.10	0	0	0	
RH	-49.30	0	0	0	-15.83	0.49	-4.21	28.77	0	0	0	*
BP	38.28	0	4.04	4.29	0.61	-4.21	0	0	0	0	8.26	
HOUR	-97.36	0	0	-3.00	-106.10	28.77	0	-196.79	0	0	0	*
noltrend	0.43	0	0	0.43	0	0	0	0	0	0	0	
Lag1	2.845	0	0	0	0	0	0	0	0	-2.85	0	*
Lag2	14.19	-1.67	0	0	0	0	8.26	0	0	0	0	

The first column indicates the main effect, next columns indicate the nonzero interactions of the row predictor and the last column indicates whether hierarchy constraint is tight or not, 0 indicate that there is no interactions between the variables. The variables with interactions are RN and Lag2, WS and BP, WindDir and BP, WindDir and HOUR,WindDir and noltrend, AirTC and BP, AirTC and HOUR, AirTC and noltrend, RH and AirTC, RH and BP, RH and HOUR, BP and Lag2.

4.3.4 GAM with pairwise interactions

The generalised additive pairwise (two-way) interactions model (GAM with the main and cross effects obtained from Table 4.3) is given by equation

(4.3.3):

$$y_t = \beta_0+\beta_1 s(\text{RN})+\beta_2 s(\text{WS})+\beta_3 s(\text{WindDir})+\beta_4 s(\text{AirTC})+\beta_5 s(\text{RH})+\beta_6 s(\text{BP}) + \beta_7 s(\text{HOUR}) + \beta_8 s(\text{noltrend}) + \beta_9 s(\text{Lag1}) + \beta_{10} s(\text{Lag2}) + \beta_{11} s(\text{RN,Lag1}) +\beta_{12} s(\text{WS,BP})+\beta_{13} s(\text{WindDir,BP})+\beta_{14} s(\text{WindDir,HOUR})+\beta_{15} s(\text{WindDir,noltrend}) +\beta_{16} s(\text{AirTC,BP})+\beta_{17} s(\text{AirTC,HOUR})+\beta_{18} s(\text{AirTC,noltrend})+\beta_{19} s(\text{RH,AirTC}) + \beta_{20} s(\text{RH,BP}) + \beta_{21} s(\text{RH,HOUR}) + \beta_{22} s(\text{BP,Lag2}) + \varepsilon_t, \quad (4.3.3)$$

where y_t is the response variable (global solar irradiance), $\beta_0, \beta_1, ..., \beta_{22}$ are the parameters and ε_t denotes the error term. The nature of the relationship between the covariates and the dependent variable is then evaluated by plotting the GAM, the results are presented in Figures 4.6 and 4.7. In Figure 4.6 the relationship between irradiance and RN, BP, noltrend, Lag1, Lag2, (WindDir and HOUR), (WindDir and Lag2) and (RH and AirTC) is linear whereas with the other variables is non-linear.

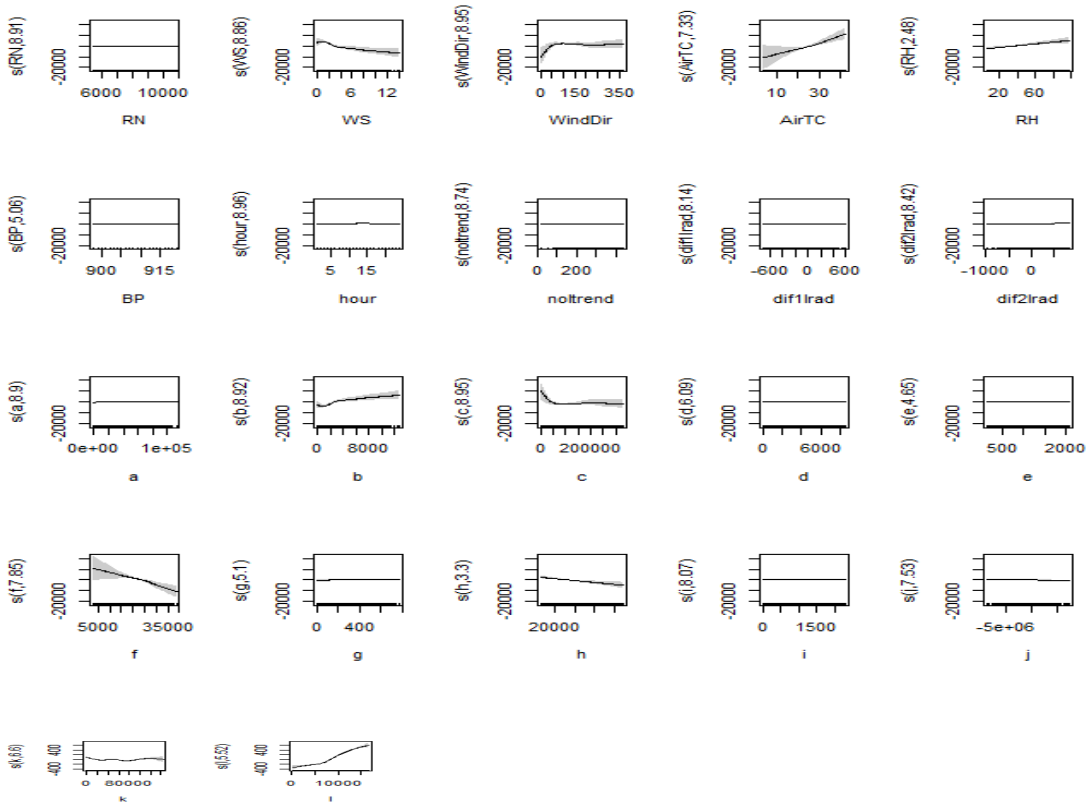

Figure 4.6: For the solar power data, the GAM with pairwise interactions model is fitted to the global solar irradiance. Each plot displays the fitted function which indicates either a linear or a nonlinear relationship.

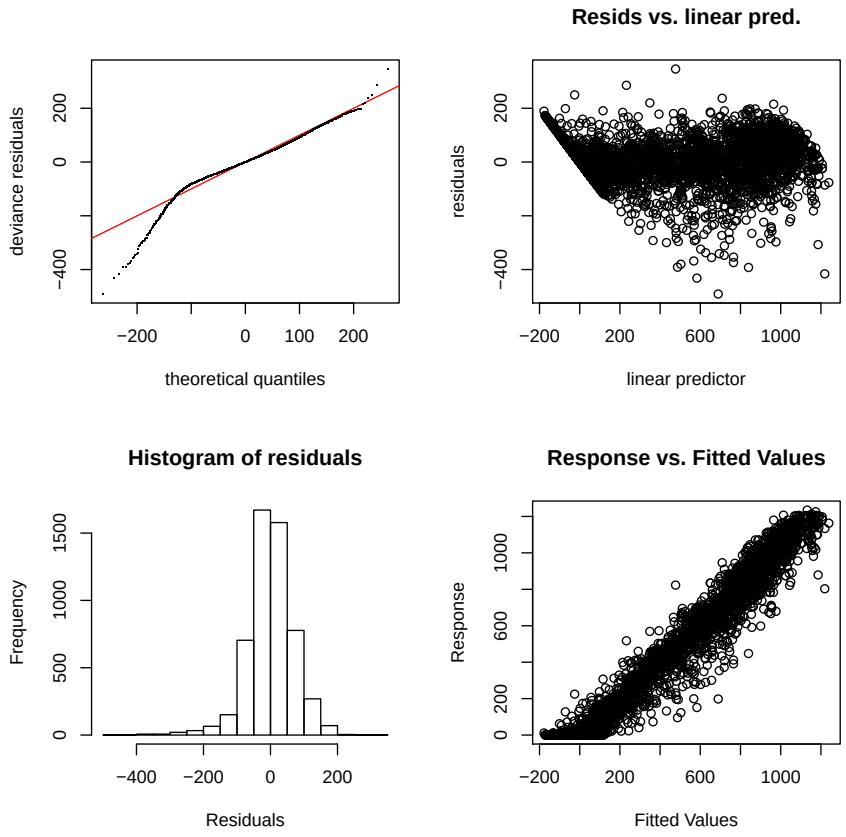

Figure 4.7: Diagnostic plots for the residuals from model 4.3.3. **Top left:** Quantile-Quantile (Q-Q) plot of residuals.

The diagnostics plots of the residuals for the GAM with the pairwise interactions are given in Figure 4.7. Top and bottom panels left plots show that the residuals are not normally distributed. Bottom panel right plot indicates that there is no sign of under and over estimation.

Residual analysis

The Ljung-Box test was applied to the residuals from the fitted GAM models (with both the main and cross effects) and got the Ljung-Box test statistic of 12269, having the p-value of $< 2.2 \times 10^{-16}$. Since the p-value for the test statistic is less than 0.05 it is concluded that there is significant autocorrelation in the residuals. To account for the residual autocorrelation, the SARIMA$(2,0,4)(1,0,1)[24]$ model given in equation 4.3.4 is considered.

$$\varepsilon_t = -y_t + \Phi_1 y_{t-24} + \phi_1 y_{t-1} - \phi_1 \Phi_1 y_{t-25} + \phi_2 y_{t-2} - \phi_2 \Phi_1 y_{t-26} \quad (4.3.4)$$

$$+\Theta_1 \varepsilon_{t-24} - \theta_1 \Theta_1 \varepsilon_{t-25} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon t - 2 - \theta_1 \Theta_1 \varepsilon_{t-26}$$

$$+\theta_3 \varepsilon t - 3 - \theta_3 \Theta_1 \varepsilon_{t-27} + \theta_4 \varepsilon_{t-4} - \theta_4 \Theta_1 + \varepsilon t - 28 \quad (4.3.5)$$

After fitting the SARIMA model and applying the Ljung-Box test the p-value of 0.6151 was obtained. Since the p-value for the test statistic is greater than 0.05 it is then concluded that the data contains no autocorrelation at 5% level of significance. The time series display of the residuals before and after accounting for the autocorrelations is given in Figure 4.8.

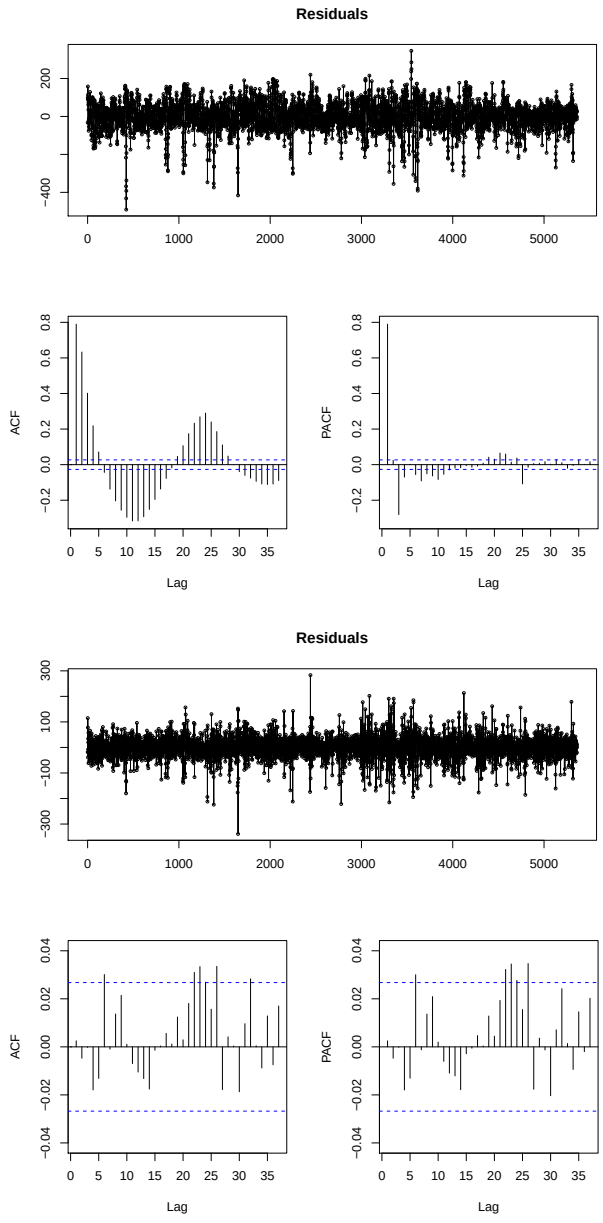

Figure 4.8: **Top panel:**Time series for the residuals before reducing autocorrelation. **Bottom panel:** Time series for the residuals after reducing autocorrelation for GAMS with interactions with a SARIMA model for model 4.3.4.

4.3.5 Model evaluation for GAM with and without the interactions coupled with SARIMA models

Forecast evaluation

The global solar irradiance is predicted with the GAM's model and compared with the forecasts generated from the model estimated on GAM with main effects only to those generated from the model estimated on the GAM with both the main and cross effects (interactions). The RMSE and MAE of M1 (GAM without interactions) are found to be 81.22 and 57.82 and for M2 (GAM with interactions) are 76.62 and 57.84 respectively. Based on the RMSE, M2 is the better model. Results for the summary of accuracy measures are shown in Table 4.4. The models for actual global solar irradiance are estimated using the training data.

Table 4.4: Accuracy measures for GAMs-SARIMA models.

Model	RMSE	MAE
Gam without interactions (M1)	81.22	57.82
Gam with interactions (M2)	76.62	57.84

Comparing the global solar irradiance and its corresponding forecasts

The forecasts distributions have been evaluated and shown in Figures 4.9 and 4.10 bottom Panels. The graphs show the actual global solar irradiance (solid line) and irradiance forecasts (dashed line) for GAMs with and without

interactions. The actual global solar irradiance and irradiance forecasts are not normally distributed and that the forecasts sometimes deviate from the true values. Top Panels illustrates the irradiance time plots. It can be seen that in Figures 4.9 and 4.10 the forecasted values (dashed line) for GAMs with and without interactions coupled with SARIMA models follow the actual global solar irradiance (solid line) remarkably well.

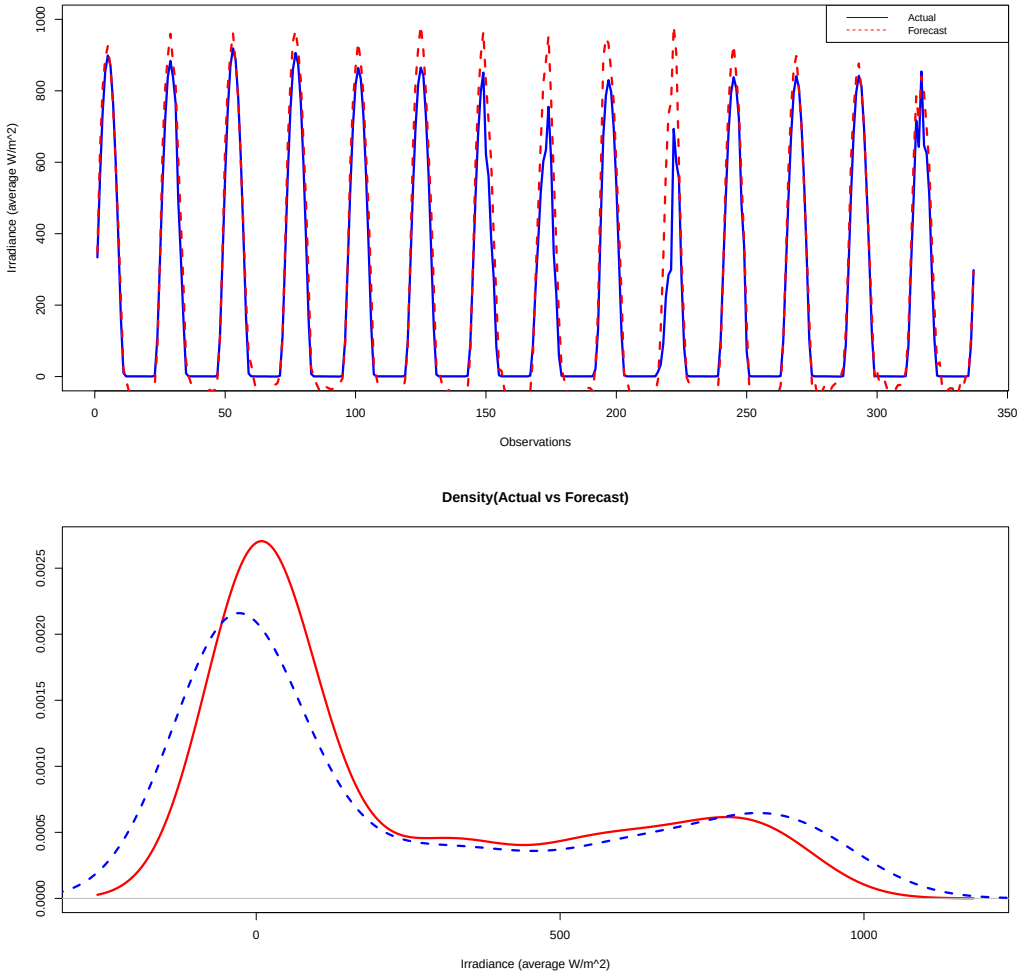

Figure 4.9: **Top panel:** Time series plot for the actual solar irradiance superimposed with its corresponding forecasts. **Bottom panel:** Time series plot for the actual solar irradiance superimposed with its corresponding forecasts for GAM without interactions coupled with SARIMA model for model 4.3.1.

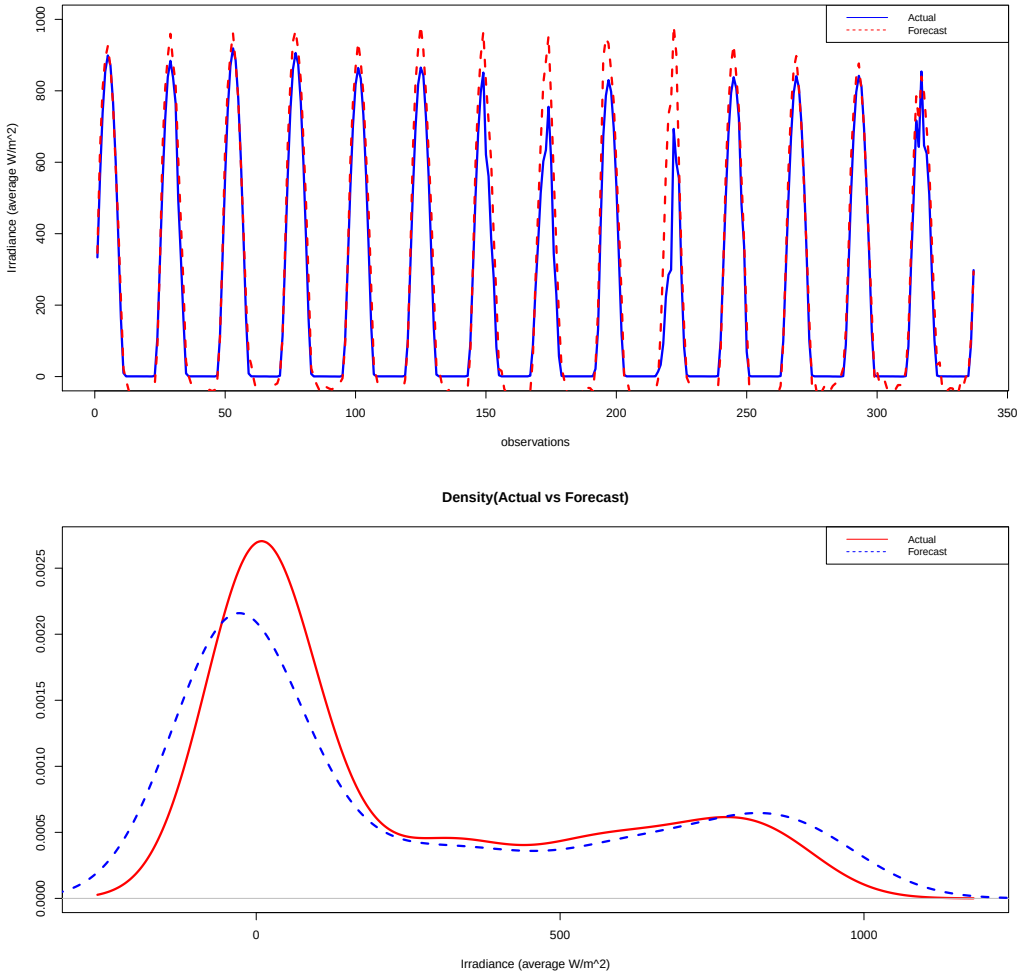

Figure 4.10: **Top panel:** Time series plot for the actual solar irradiance superimposed with its corresponding forecasts. **Bottom panel:** Time series plot for the actual solar irradiance superimposed with its corresponding forecasts for GAM with interactions coupled with SARIMA model for model 4.3.3.

4.3.6 Model selection for GAMS with and without the interactions with SARIMA models

The best model is selected based on the $AdjR^2$, AIC, BIC and GCV. The preferred model is the one with the minimum value of AIC, BIC and GCV and with the maximum $AdjR^2$ value compared to the other models. The GAM with the pairwise interactions (model M2) is considered to be the better model with an $AdjR^2$ value of 0.967, AIC of 60968.25, BIC of 61899.98 and finally a GCV of 5089 (Table 4.5).

Table 4.5: Model selection for GAM with and without interactions with SARIMA model.

Models	$AdjR^2$	AIC	BIC	GCV
GAM without interactions (M1)	0.952	62903.14	63318.5	7298
GAM with interactions (M2)	0.967	60968.25	61899.98	5089

4.4 Partially linear additive quantile regression models

This section discusses two PLAQR models, the one with and without interactions. PLAQR without interaction is the one containing all the predictor variables and the PLAQR model with interactions is the one which includes all the significant variables selected using Lasso via hierarchical interactions.

4.4.1 Partially linear additive quantile regression models with and without interactions

The PLAQR model considers both linear and nonlinear covariates. The PLAQR models with and without interactions are given in equations 4.4.1 and 4.4.2, respectively:

$$y_{t1} = \beta_0+\beta_1(\text{RH})+\beta_2(\text{BP})+bs(\text{RN})+bs(\text{WS})+bs(\text{WindDir})+bs(\text{AirTC})+bs(\text{HOUR}) + bs(\text{noltrend}) + bs(\text{Lag1}) + bs(\text{Lag2}) + \varepsilon_t \quad (4.4.1)$$

and

$$\begin{aligned} y_{t2} = \beta_0 &+ \beta_1(\text{RN}) + \beta_2(\text{RH}) + \beta_3(\text{BP}) + \beta_4(\text{noltrend}) + \beta_5(\text{Lag1}) \\ &+ \beta_6(\text{Lag2}) + \beta_6(\text{WindDir,HOUR}) + \beta_7(\text{WindDir,noltrend}) \\ &+ \beta_8(\text{RH,AirTC}) + bs(\text{WS}) + bs(\text{WindDir}) + bs(\text{AirTC}) \\ &+ bs(\text{HOUR}) + bs(\text{RN,Lag2}) + bs(\text{WS,BP}) + bs(\text{WindDir,BP}) \\ &+ bs(\text{AirTC,BP}) + bs(\text{AirTC,HOUR}) + bs(\text{BP,Lag2}) + \varepsilon_t, \quad (4.4.2) \end{aligned}$$

where bs is a B-spline. The interactions presented in the above equations were found after doing variable selection using the Lasso via hierarchical interactions, and the variables selected are RN and Lag2, WS and BP, WindDir and BP, WindDir and HOUR,WindDir and noltrend, AirTC and BP, AirTC and HOUR, AirTC and noltrend, RH and AirTC, RH and BP, RH and HOUR, BP and Lag2.

4.4.2 Residual analysis for PLAQR with and without interactions models

After fitting the PLAQR models a test for autocorrelation of the residuals is carried out using the Ljung-Box test and found a p-value of $< 2.2 \times 10^{-16}$ for both PLAQR models with and without interaction. Since the p-values are less than 0.05 it is concluded that the data contain significant autocorrelation. In order to account for these autocorrelations for the PLAQR model without interactions, the SARIMA$(4,1,4)(2,0,5)[24]$ model is fitted. This is followed by testing for the autocorrelations using Ljung Box-test and found the p-value to be 0.3974 ($p > 0.05$). For the PLAQR model with the pairwise interactions the SARIMA$(4,1,4)(2,0,3)[24]$ is fitted to the residuals and then tested for autocorrelation and obtained a p-value of 0.2809 ($p > 0.05$). Figures 4.11 and 4.12 give the time series plots for PLAQR models residuals with and without the interactions coupled with SARIMA models.

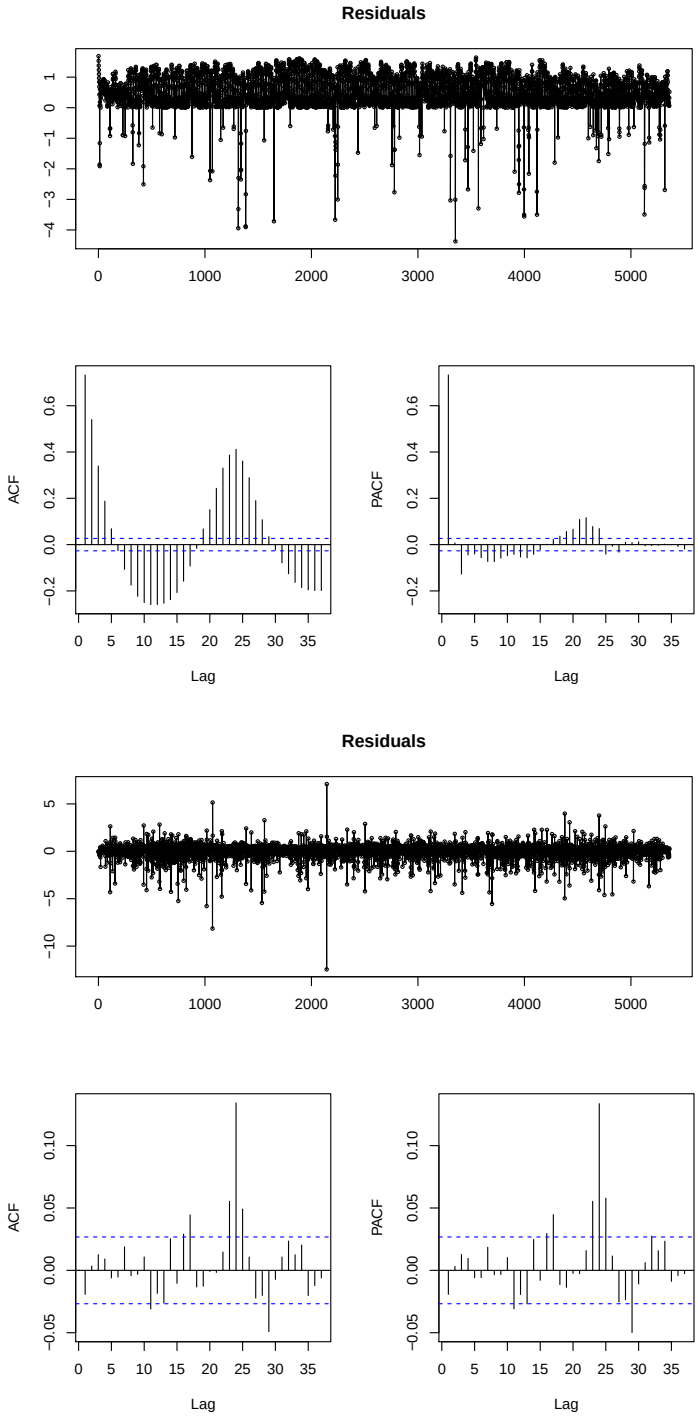

Figure 4.11: **Top panel:**Time series for the residuals before reducing autocorrelation.**Bottom panel:**Time series for the residuals after reducing autocorrelation for PLAQR model without the interaction coupled with SARIMA models for model 4.4.1.

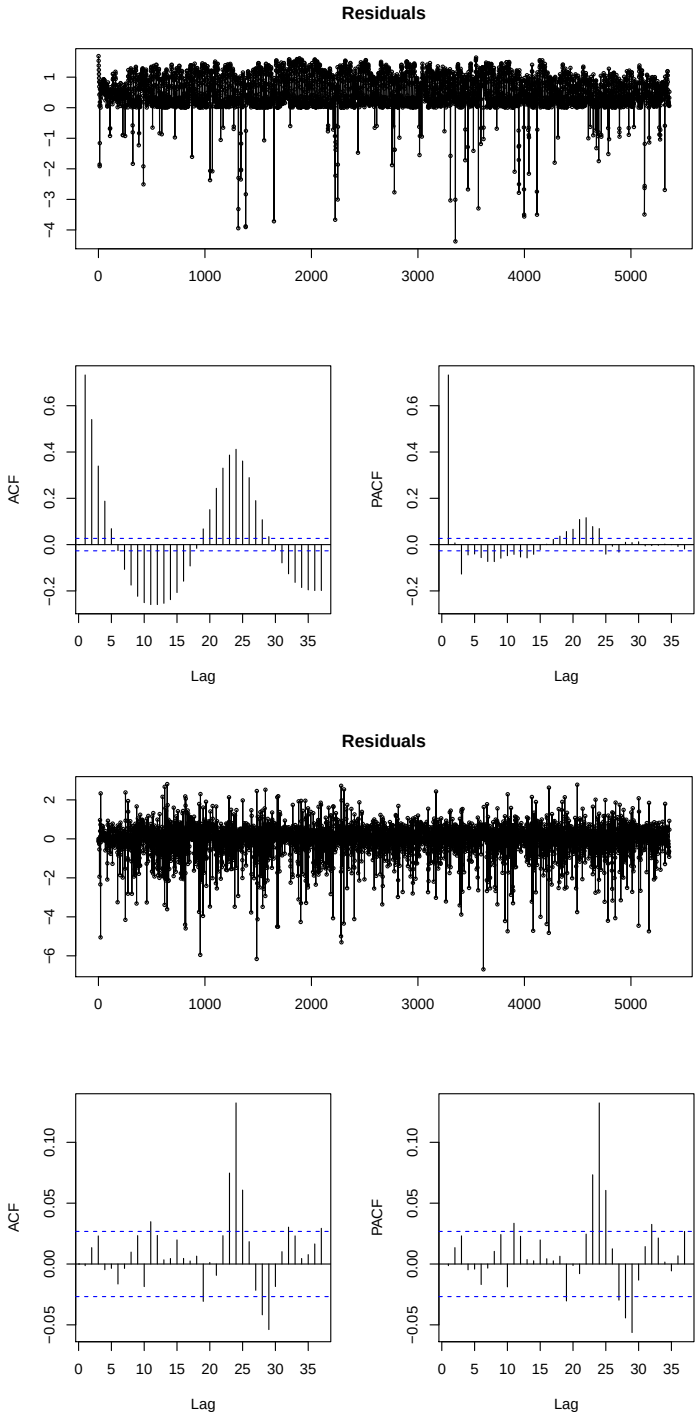

Figure 4.12: **Top panel:**Time series for the residuals before reducing autocorrelation. **Bottom panel:**Time series for the residuals after reducing autocorrelation for PLAQR model with interaction with SARIMA models for model 4.4.2.

4.4.3 Model evaluation for PLAQR with and without pairwise interactions with SARIMA models

Forecast evaluation

Table 4.6 gives the summary of the forecast accuracy measures for PLAQR models without interactions (M3) and with interactions (M4). The value of RMSE is found to be 96.26 for model M3 and 83.29 for model M4. The value of MAE is 60.07 for model M3 and 52.34 for model M4. This means that model M4 is the better model given that its RMSE and MAE are found to be small in comparison with those of model M3.

Table 4.6: Forecast evaluation for PLAQR with and without interaction coupled with SARIMA models.

Models	RMSE	MAE
PLAQR without interactions (M3)	96.26	60.07
PLAQR with interactions (M4)	83.29	52.34

Comparing the actual solar irradiance and its corresponding forecasts

It can be seen that in Figures 4.13 and 4.14, top panel the forecast values (dashed line) for PLAQR models with and without interactions follow the actual irradiance (solid line) remarkably well. There is an over estimation in Figure 4.13 between observations 0 and 300 and in Figure 4.14 between observations 150 and 350. In the bottom panel, the graphs show densities of the actual solar irradiance (solid line) and irradiance forecasts (dashed line). It can be seen that actual solar irradiance and irradiance forecast are not normally distributed.

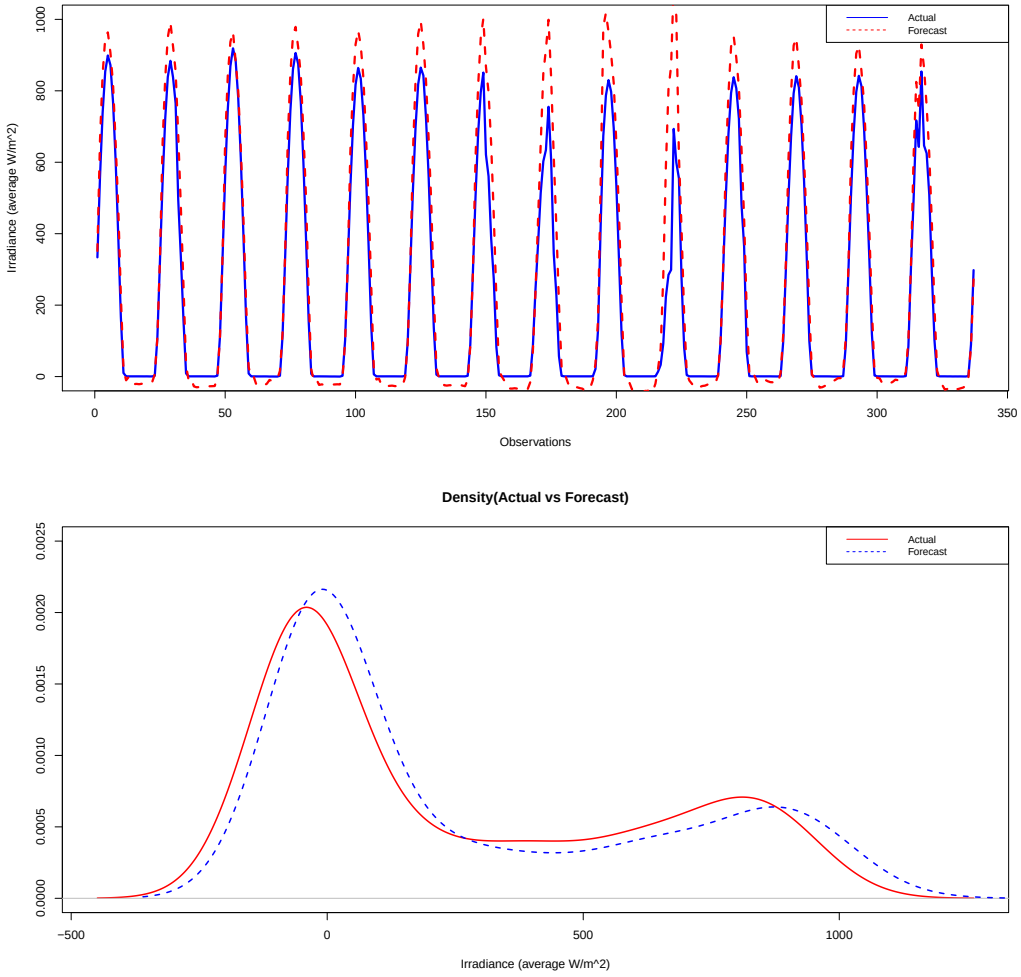

Figure 4.13: **Top panel:** Time series plots for global solar irradiance superimposed with global solar irradiance forecasts. **Bottom panel:** Density plots of solar irradiance superimposed with irradiance for PLAQR without interaction coupled with SARIMA models for model 4.4.1.

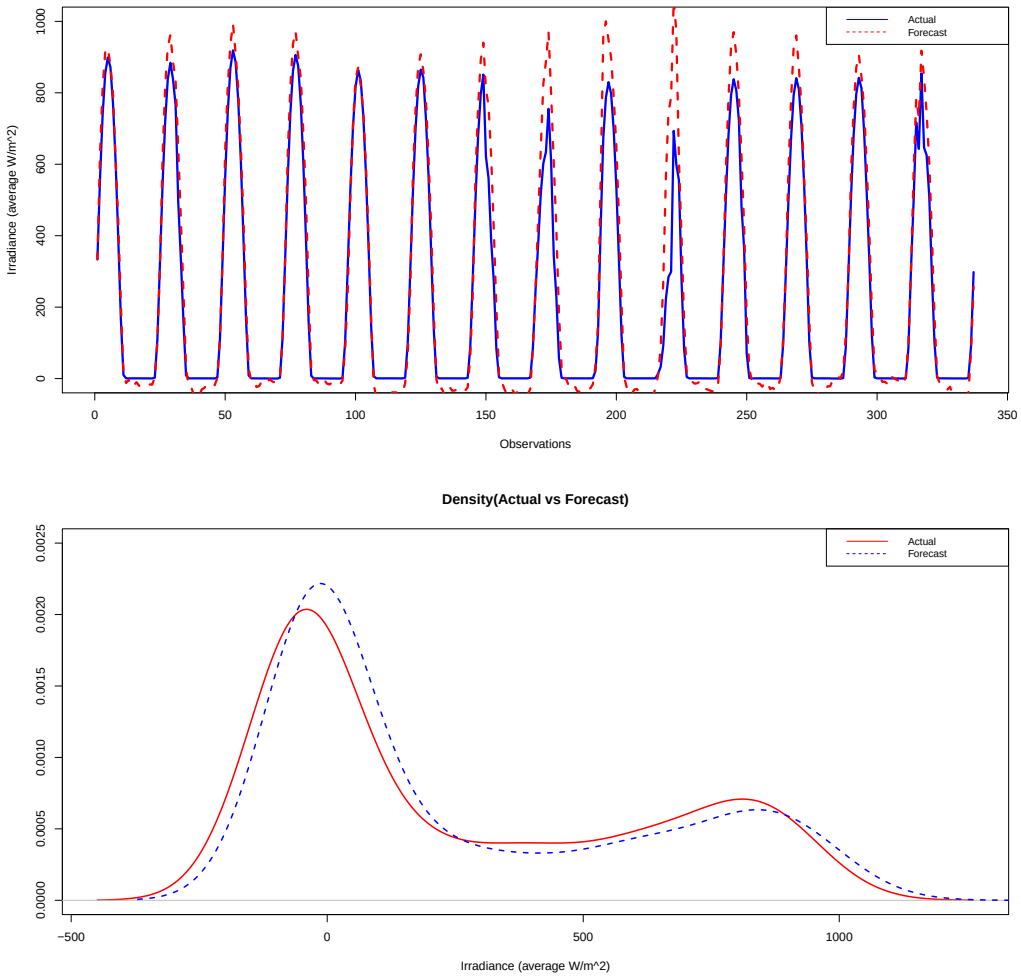

Figure 4.14: **Top panel:** Time series plots for global solar irradiance superimposed with global solar irradiance forecasts. **Bottom panel:** Density plots of solar irradiance superimposed with irradiance for PLAQR with interaction coupled with SARIMA model for model 4.4.2.

4.4.4 Model selection for PLAQR with and without the interactions with SARIMA models

The best model is selected using AIC, BIC and the $AdjR^2$. The preferred model is the one with the minimum value of AIC, BIC and the maximum value of $AdjR^2$ compared to the other models. The model with the pairwise interactions (model M4) is considered to be the best model with the $AdjR^2$ value of 0.957, AIC of 60056.53 and finally the BIC of 60744.76 (Table 4.7).

Table 4.7: Model selection for PLAQR with and without interactions.

Models	$AdjR^2$	AIC	BIC
PLAQR(M3)	0.947	61191.64	60748.55
PLAQR with interactions (M4)	0.954	60744.76	60056.53

4.5 Forecast combination

The "opera" r-package developed by (Gaillard, 2015) is used to combine the forecasts from the best models, which are models M2 and M4, respectively. This is done in order to improve the accuracy of forecasts obtained from these models. The convex combination of the models was done using an algorithm based on the pinball, the absolute, the percentage error and the square losses. Based on the pinball loss function the average loss suffered by the PLAQR model with pairwise interactions (model M4) is smaller than that of the GAM with pairwise interactions (model M2) as seen in Figure

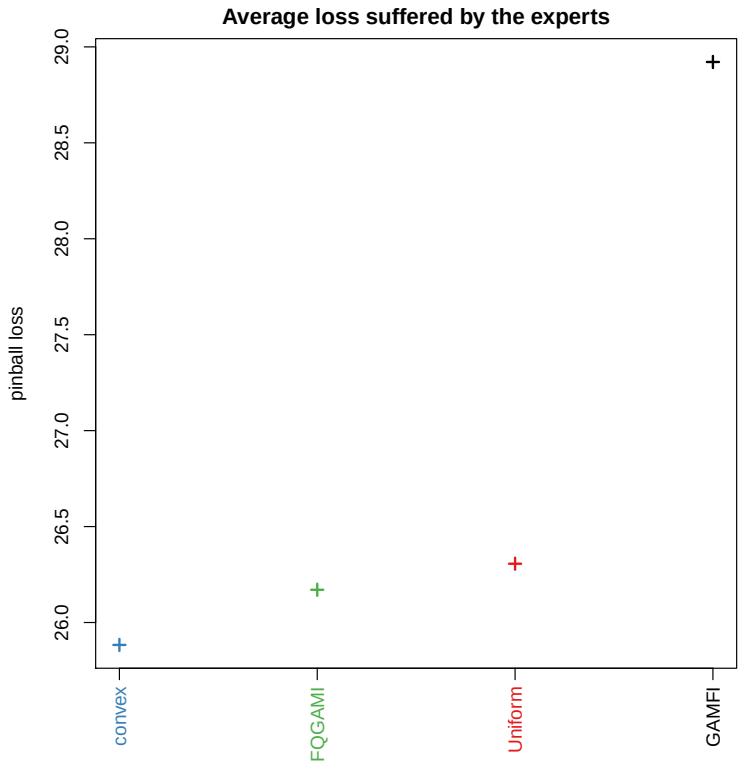

Figure 4.15: Average loss suffered by the models.

4.15, where GAMFI and CONVEX are the probabilistic global solar irradiance forecasts for GAM with interaction and the CONVEX combination models, respectively. This means that the PLAQR model with pairwise interactions produce more accurate forecasts compared to the GAM with pairwise interaction.

The weights assigned to the forecasts for model M2 and model M4 are found to be 0.18 and 0.82 respectively. Based on the pinball loss, model M2 is considered to be a better model. Figure 4.16 gives a summary of the models, the accuracy measures for the parameter loss type and the average pinball.

Models (experts)		
	M2	M4
Pinball loss (Mix1)	0.18	0.82
Absolute error (Mix2)	0.18	0.82
Percentage error (Mix3)	5.97×10^{-21}	1
Square loss (Mix4)	0.668	0.332

Accuracy measures				
	Mix1	Mix2	Mix3	Mix4
RMSE	79.23	79.23	83.29	79.62
MAE	51.77	51.77	52.34	58.50

Average pinball losses mean			
	M2	M4	Combined
Pinball loss	28.92	26.17	25.88

Figure 4.16: Evaluation of forecast combination models.

Table 4.8 gives a summary of the accuracy measures for the models M2 and M4 along with the combined forecasts (M5). Based on MAE, model M4 is the best compared to model M2. There is a slight improvement on the accuracy measures after forecast combination.

Table 4.8: Model comparisons of the best GAM and best PLAQR and combined.

Accuracy measures	Model M2	Model M4	Combined (M5)
RMSE	76.62	83.29	79.23
MAE	57.84	52.34	51.77

A plot of irradiance and the combined forecasts is given in Figure 4.17 top panel, and indicates that the forecasts follow the actual irradiance very well. The densities between the actual irradiance and the combined forecasts are given in Figure 4.17 bottom panel.

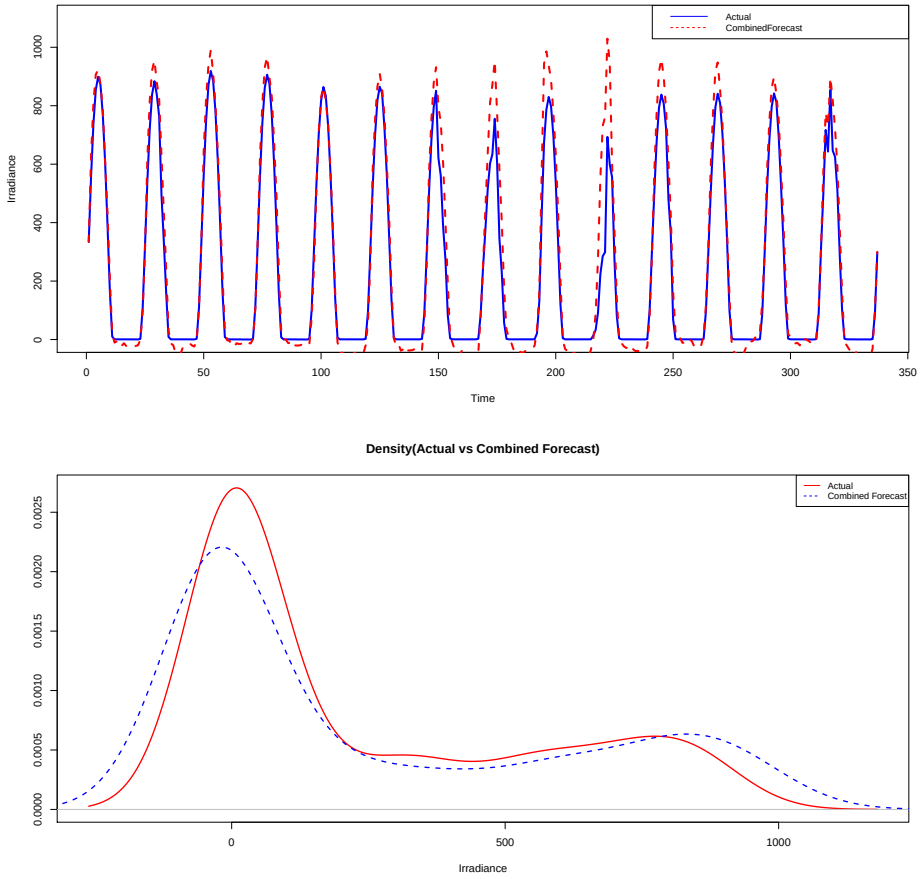

Figure 4.17: **Top panel:** Plot of Actual irradiance superimposed with the combined forecasts. **Bottom panel:** Density plot of Actual irradiance superimposed with the combined forecasts.

4.6 Quantile regression averaging of individual probabilistic forecasts

In order to combine forecasts and compute the prediction intervals for the developed models the quantile regression averaging (QRA) is used. QRA

was first introduced by Nowotarski and Weron (2015). Quantile regression is applied to the probabilistic forecasts of the forecasting models. Probabilistic solar irradiance forecasts are used as independent variables and solar irradiance as the dependent variable. The generalised additive model with interaction (Model M2) forecasts and the convex (Model M5) model forecasts are regressed with the actual solar irradiance data. The models are presented by equations (4.6.1) and (4.6.2), respectively.

$$y_{t,\tau(\text{GAMA})} = \beta_{0,\tau} + \beta_{1,\tau}\text{GAMFI} + \varepsilon_{t,\tau} \tag{4.6.1}$$

$$y_{t,\tau(\text{CONVEXA})} = \beta_{0,\tau} + \beta_{1,\tau}\text{CONVEX} + \varepsilon_{t,\tau}, \tag{4.6.2}$$

where GAMA is the GAM averaging, CONVEXA is the CONVEX averaging. *Note:* Equations (4.6.1) and (4.6.2) are denoted by model M6 and model M7, respectively.

4.6.1 Comparing the performance of the GAM and PLAQR with interactions with the GAM and PLAQR Averaging

Table 4.9 gives a summary of the accuracy measures for the models M2 and M5 along with their corresponding averaging model. It can be seen that there is a significant change on the accuracy measures. The RMSE and MAE for Model M2 have changed from 76.62 to 45.12 and from 57.84 to 20.37. The RMSE and MAE for model M5 have changed from 79.73 to 49.21 and from 51.77 to 19.46.

Table 4.9: Comparison of the GAM and PLAQR before and after regression.

Before regressing		
Accuracy measure	Model M2	Model M5
RMSE	76.62	79.23
MAE	57.84	51.77
After regressing (Averaging models)		
Accuracy measure	Model M6	Model M7
RMSE	45.13	49.21
MAE	20.37	19.46

4.6.2 Quantile regression averaging

QRA is an extension of combining probabilistic forecasts and is given by the following equation:

$$y_{t,\tau(\text{QRA})} = \beta_{0,\tau} + \beta_1 \text{GAMQRA} + \beta_{2,\tau}\text{PLAQRA} + \beta_{3,\tau}\text{CONVEQRA} + \varepsilon_{t,\tau}, \tag{4.6.3}$$

where GAMQRA, PLAQRA and CONVEXQRA are the regressed GAM, PLAQR and CONVEX forecasts, in this case taken as predictors of solar irradiance.

Residual analysis for the QRA

The Ljung BOX-test is applied to the residuals up to lag 15 from the fitted QRA model and a Ljung Box-test statistic of 2320.2 is obtained, with the p-value of less than 2.2×10^{-16}. Since the p-value for the test statistic regarding the QRA is less than 0.05, it is concluded that the solar irradiance

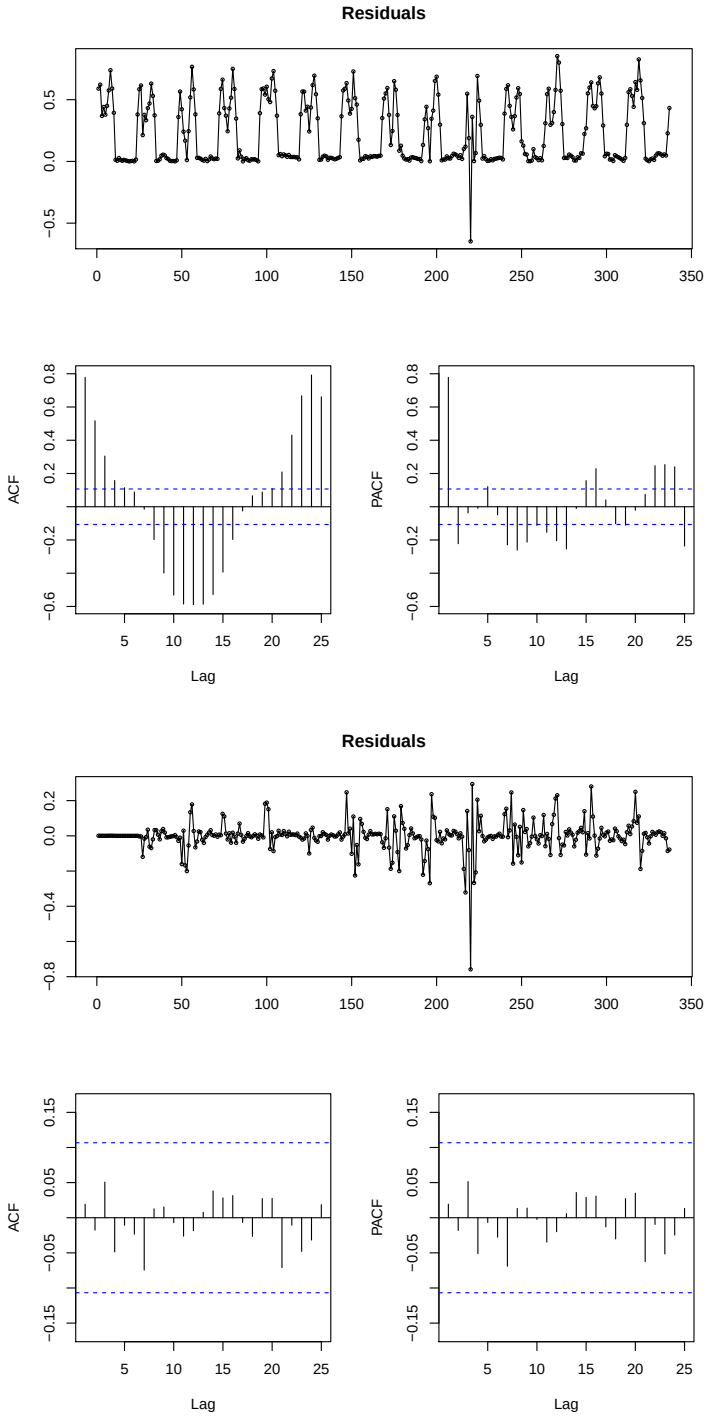

Figure 4.18: **Top panel:** Time series display of residuals of QRA before correcting for autocorrelation. **Bottom panel:** Time series display of residuals of QRA after correcting for autocorrelation for model 4.6.3 coupled with a SARIMA model.

data has significant autocorrelation. In order to reduce the autocorrelation, SARIMA$(3,0,2)(1,1,1)[24]$ is fitted to the residuals and a p-value of 0.9847 is obtained. Since the p value obtained is greater than 0.05it is concluded that the residuals are not autocorrelated at the 5% level of significance. Time series display of residuals analysis of QRA before and after correcting it for autocorrelation are given in Figure 4.18.

Comparing the actual solar irradiance and its corresponding forecasts

The value of RMSE for the QRA is found to be 41.33 and its MAE to be 17.64. The density plot of the forecasts of solar irradiance superimposed with the actual solar irradiance is given in the bottom panel of Figure 4.19. The plot (Figure 4.19) suggests that actual solar irradiance and irradiance forecasts are not normally distributed. Top Panel illustrates the irradiance time series plots. In both panels in Figure 4.19 the forecasted values (dashed lines) for QRA follow the actual irradiance (solid line) remarkably well.

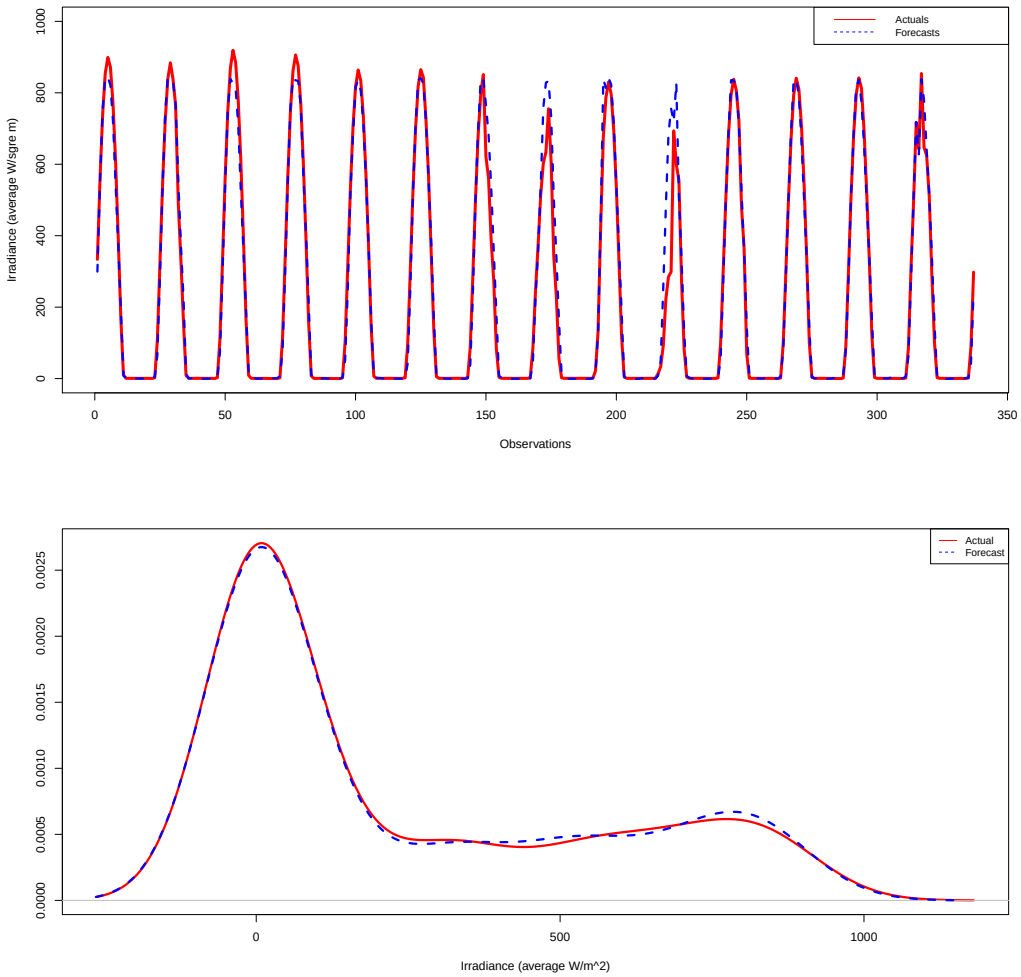

Figure 4.19: **Top panel:** Time series display for the actual solar irradiance and its corresponding forecasts for model (4.6.3). **Bottom panel:** Density plot for the actual solar irradiance and its corresponding forecasts with SARIMA models for model (4.6.3).

4.7 Comparative analysis of the models

This section presents the comparison of the best models (Model M4, M6, M7 and M8) using prediction intervals and forecast error distributions, and also the improvement rates of the best over the other models.

4.7.1 Prediction interval widths

From Table 4.10 the PINAW of model M8 is lower than that of models M4, M6 and M7. This implies that model M8 produces a narrower PINAW, therefore model M8 is the best compared to models M4, M6 and M7. The PICP for model M8 is 99.41%, which is valid since its greater than 95%. This indicates that the prediction intervals constructed by model M8 are more informative with a valid PICP ($\geq$ 95%). The PICP is set to 95% level of confidence and is valid for all models. Model M8 also gives smallest PINAD compared to the other models. This confirms that M8 is the best model.

Table 4.10: Comparison of the best models using the PICP, PINAW and PINAD at 95% level of confidence.

Models	PICP	PINAW	PINAD
Model M4	98.52%	27.01%	0.08%
Model M6	98.22%	16.22%	0.05%
Model M7	97.92%	16.94%	0.05%
Model M8	99.41%	12.24%	0.03%

Figure 4.20 displays box-plots of the models and the density plots which indicates that none of the models had a good coverage based on 95% PI since they have the PICP that is less than 100% . In Figure 4.20, bottom panel it is also observed that in terms of prediction intervals, model M8 outperforms the other models; that is it selected M8 as the best forecasting model.

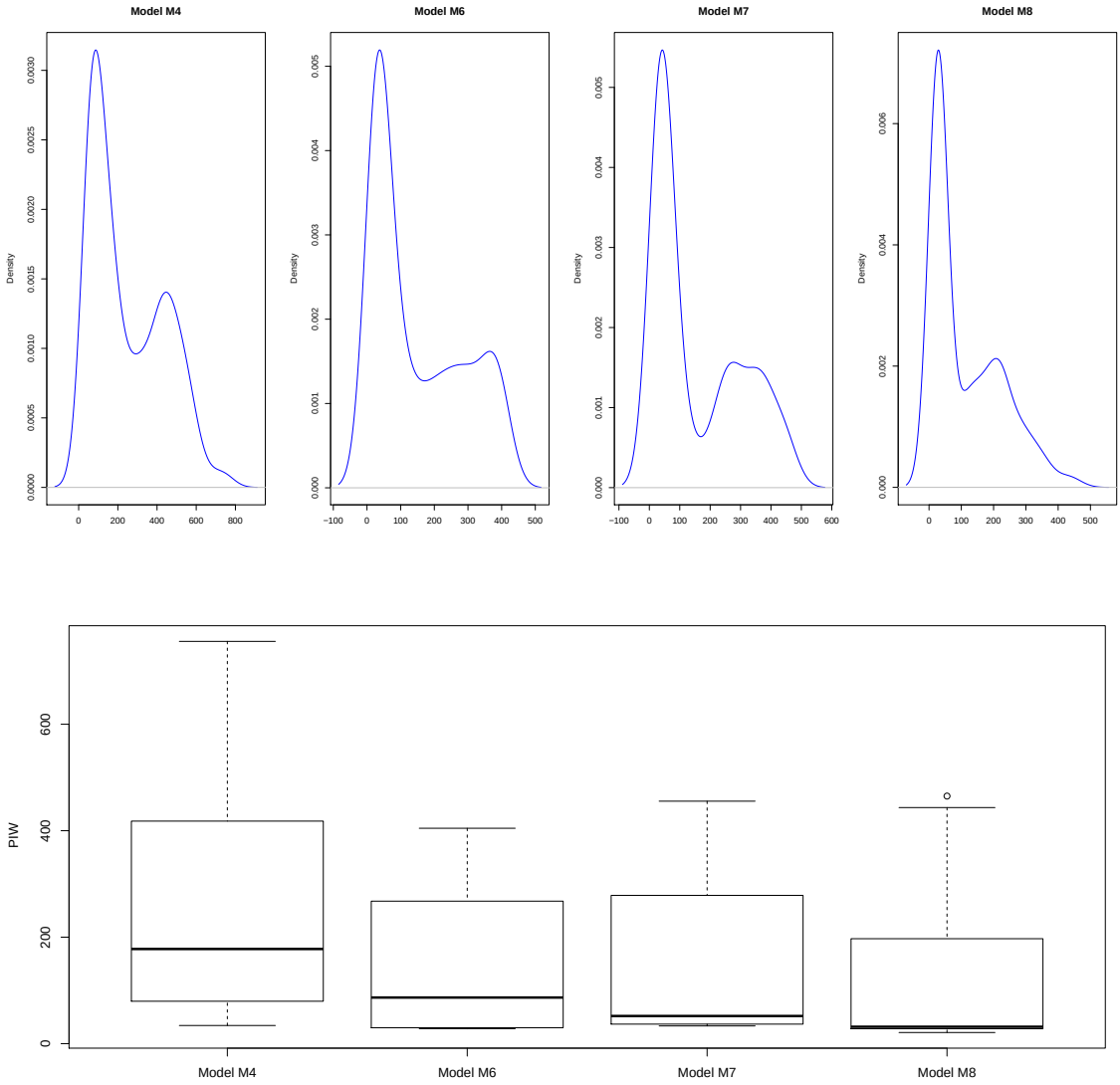

Figure 4.20: **Top panel:** Density plots of the PIWs for models M4, M6, M7 and M8, respectively. **Bottom panel:** Box plots of the PIWs for models M4, M6, M7 and M8, respectively.

4.7.2 Forecast error distributions

Table 4.11 gives summary statistics of the residuals of the models. Comparing solar irradiance forecasts with the observed values, the mean forecast error for model M8 is the lowest compared to that of models M4, M6 and M7 when under predicting the solar irradiance. For over predictions model M4 has a mean of 0.45 which is the lowest compared to models M6 (0.57), M7 (0.57) and M8 (0.58). This is compared to the frequency of over predictions and under predictions. Model M6 has the smallest standard deviation, this implies that M6 is the best compared to other models. A negative skewness in the distribution reflects a large number of negative errors, which reflects the overestimation of solar irradiance. Such a pattern can be seen in Figure 4.21 (top panel) where large negative residuals dominate positive ones. The distributions of the errors are highlighted further by the box and whisker plots in Figure 4.21 (bottom panel).

Table 4.11: Summary statistics of the residuals of the models.

Summary statistics of residuals				
	Model M6	Model M4	Model M7	Model M8
Standard deviation	0.67	79.53	48.85	41.14
Mean	-0.56	-25.12	-6.56	-4.50
Median	-0.13	4.49	-0.10	-0.11
1st Qu	-1.04	-66.21	-0.70	-0.94
3rd Qu	-0.05	26.23	1.69	1.57
Skewness	-1.34	-2.34	-4.04	-3.45
Kurtosis	1.61	7.89	24.46	21.25
Under and over predictions				
Mean under predicted	0.43	0.53	0.43	0.42
Mean Over predicted	0.57	0.47	0.57	0.58
Under prediction	144	179	145	141
Over prediction	193	158	192	196

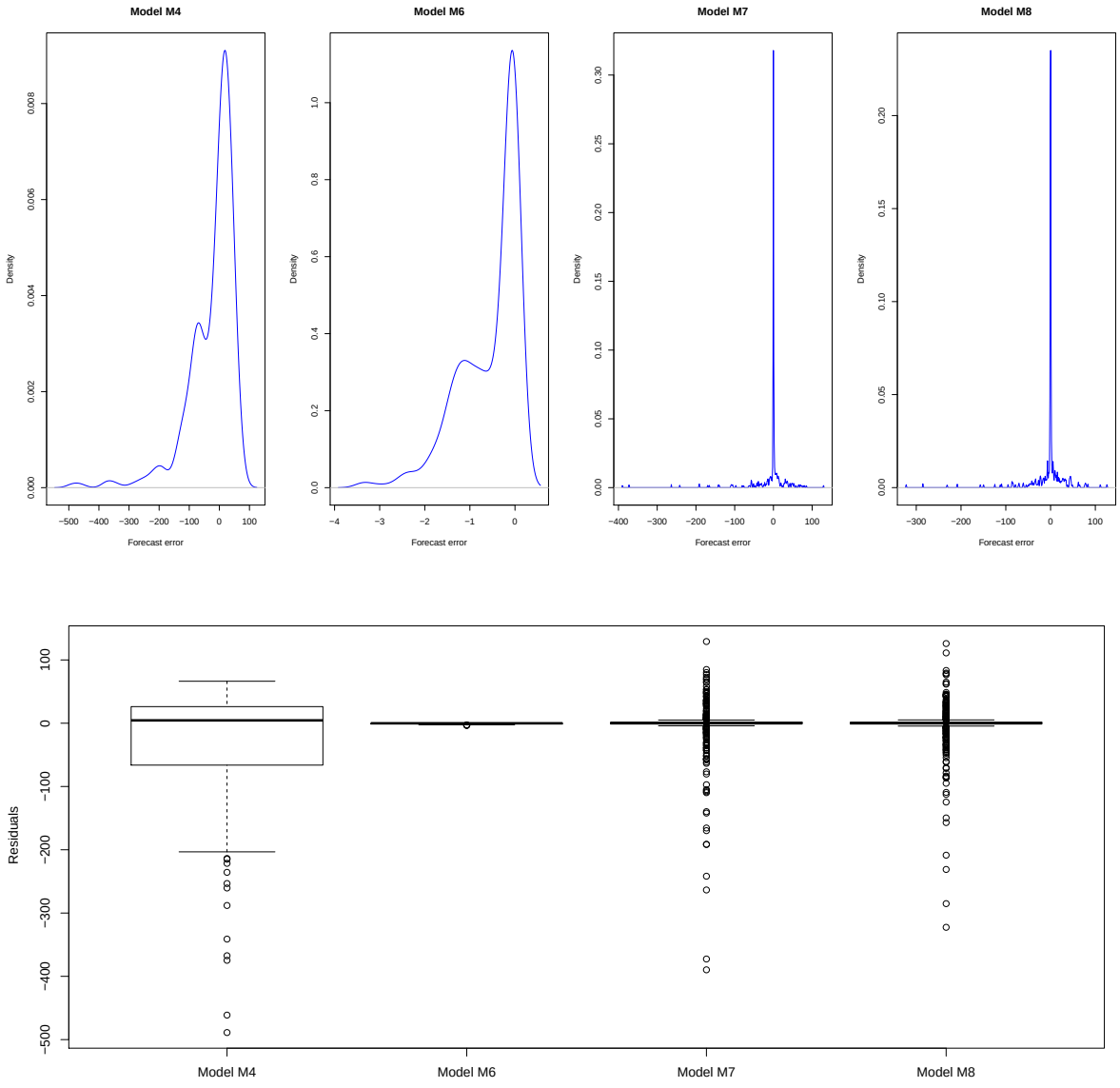

Figure 4.21: **Top panel:** The error distributions for models M4, M6, M7 and M8, respectively. **Bottom panel:** Box and whisker plots for models M4, M6, M7 and M8, respectively.

4.7.3 Percentage improvement

Comparison of the best models using the PICP, PINAW and PINAD is done and the results are presented in Table 4.10. Based on the PINAW, M8 is found to be the best model. The performance of the models against the best models which is model M8 is then calculated and results are presented in Table 4.12. The percentage improvements of M8 over M4, M6 and M7 are found to be 13.87%, 8.49% and 18.98%, respectively.

Table 4.12: Percentage improvement rates of the best model over other models.

Best model over other	Percentage improvement
Model M8 over M4	13.87%
Model M8 over M6	8.49%
Model M8 over M7	18.98%

Chapter 5

Conclusion

5.1 Introduction

This dissertation discussed an application of partially linear additive quantile regression (PLAQR) models to probabilistic solar power forecasting using data from Tellerie radiometric station in South African. The models were developed based on hourly global solar irradiance data from Eskom, South Africa's power utility company. The fitted models were compared with generalised additive models (GAMs) and forecast combination models, namely: the convex combination and quantile regression averaging (QRA) models.

5.2 Summary

Chapter 1 discussed the main aim of the study which was to develop probabilistic medium-term solar power forecasting models for predicting solar power generation for some of the solar plants located in South Africa. The

chapter also discussed the main objectives of the study which were to: develop PLAQR models for probabilistic solar power forecasting, estimate the parameters of the developed models using the Barrodale and Roberts algorithm for ℓ_1-regression, evaluate the performance of the models, forecast the hourly global solar irradiance and evaluate the accuracy of the forecasts.

Chapter 2 provided the summaries of some studies that used the proposed methodology to forecast global solar irradiance. It is concluded that the use of PLAQR models and GAMs on probabilistic solar power forecasting has not been done to date in South Africa. Chapter 3 discussed the general theory of the proposed methods, which were GAMs, quantile regression (QR) and PLAQR models.

Chapter 4 presented a discussion of the empirical results. Section 4.3 looked at an application of GAMS. Firstly the GAM without interactions given by equation 4.3.1 was considered and a SARIMA model given by equation 4.3.2 was used to reduce the autocorrelation. Secondly the GAM with interactions given by equation 4.3.3 was discussed. The interactions were obtained using a shrinkage method called Lasso via hierarchical interactions. The performance of the models was evaluated using the accuracy measures root mean square error (RMSE) and mean absolute error (MAE), and the best models were selected based on the Bayesian information criterion (BIC), Akaike information criterion (AIC), adjusted R-squared ($AdjR^2$) and generalised cross validation (GCV). The results illustrated that the GAM with interactions is the best model, with the BIC value of 61899.98, AIC of 60968.25, $AdjR^2$ of 0.967 and GCV of 5089.

In Section 4.4 an application of (PLAQR) models was presented. A PLAQR model is a combination of GAMS and QR models and it is used to relax the assumption that there is a linear relationship between the covariates in quantile regression. The performance of the models were evaluated using the accuracy measures RMSE and MAE and the best model was selected based on the BIC, AIC, $AdjR^2$ and GCV. The results illustrated that the PLAQR model with interactions is the best model, with BIC value of 60056.53, AIC of 60744.76, $AdjR^2$ of 0.954.

The forecast combination of the two best models, the GAM with interactions and PLAQR with interactions were presented in Section 4.5. The main purpose on this section was to improve the accuracy measures obtained by these two models. The results showed that, based on MAE the PLAQR with interactions gives better forecasts compared to the GAM with interactions and that there is a slight improvement on the accuracy measures after forecast combination.

The forecast combination was then extended in Section 4.6 to QRA, this resulted in four sets of forecasts. To compute the prediction intervals for the GAM with interaction and the CONVEX model (obtained from forecast combination, denoted by Model M5) the QRA was used. The solar irradiance was regressed on the forecasts generated by the GAM and the CONVEX models resulting in models M6 (GAMA) and M7 (CONVEXA), respectively. This resulted in an improvement of the forecasts. The GAMA and CONVEXA models gave more accurate forecasts compared to the individual GAM and CONVEX models. Therefore the best set of forecasts was selected based

on the prediction interval coverage probability (PICP), prediction interval normalised average width (PINAW) and prediction interval normalised average deviation (PINAD) given in Section 4.7.1. The results illustrated that QRA is the best since it produces robust prediction intervals as compared to the other models. The results then showed that QRA (Model M8) was the best model with narrowest PINAW=12.24% and smallest PINAD=0.03% and with valid PICP ($> 95\%$).

Finally in Section 4.7.3 the percentage improvements between the best model and the other models were calculated. The results showed that model M8 yields a small improvement over model M6, whereas M8 yields higher percentage improvement over M7.

5.3 Concluding remarks

- The PLAQR models for probabilistic solar power forecasting were developed in this dissertation.
- The parameters of the developed models in this dissertation were estimated using the Barrodale and Roberts algorithm for ℓ_1-regression.
- The performance and the accuracy of the forecasts were evaluated and the PLAQR model with pairwise interactions was found to be an appropriate model for forecasting hourly solar irradiance.

5.4 Contributions

The main contribution of this dissertation is the inclusion of non-linear trend variables and the extension of combining forecasting models using QRA. This study could be useful to system operators, including decision-makers in power utility companies such as Eskom.

5.5 Limitations of the dissertation

There are several predictors of solar irradiance that are well known. This dissertation is limited to conclude about the global solar irradiance only because forecasting of GHI is the most essential step in most PV power prediction systems. The data used in the study provided only few predictor variables. These were the only available variables. In this dissertation the analysis was based on summer because the data used was recorded during summer season.

5.6 Future research

PLAQR models assume that some covariates have a linear relationship with the response while other covariates have an unknown non-linear relationship. It also assumes additive form between the non-linear covariates, which allows

for a flexible model. In this dissertation the PLAQR models were modelled using B-splines so as to capture the non-linear relationships between some of the covariates with the response variable. An application of PLAQR models using two different datasets, but of the same sample size and an inclusion of more covariates would be important for further research. Future research could also look at developing models for forecasting global solar irradiance for each of the four seasons of the year.

Appendices

Appendix A

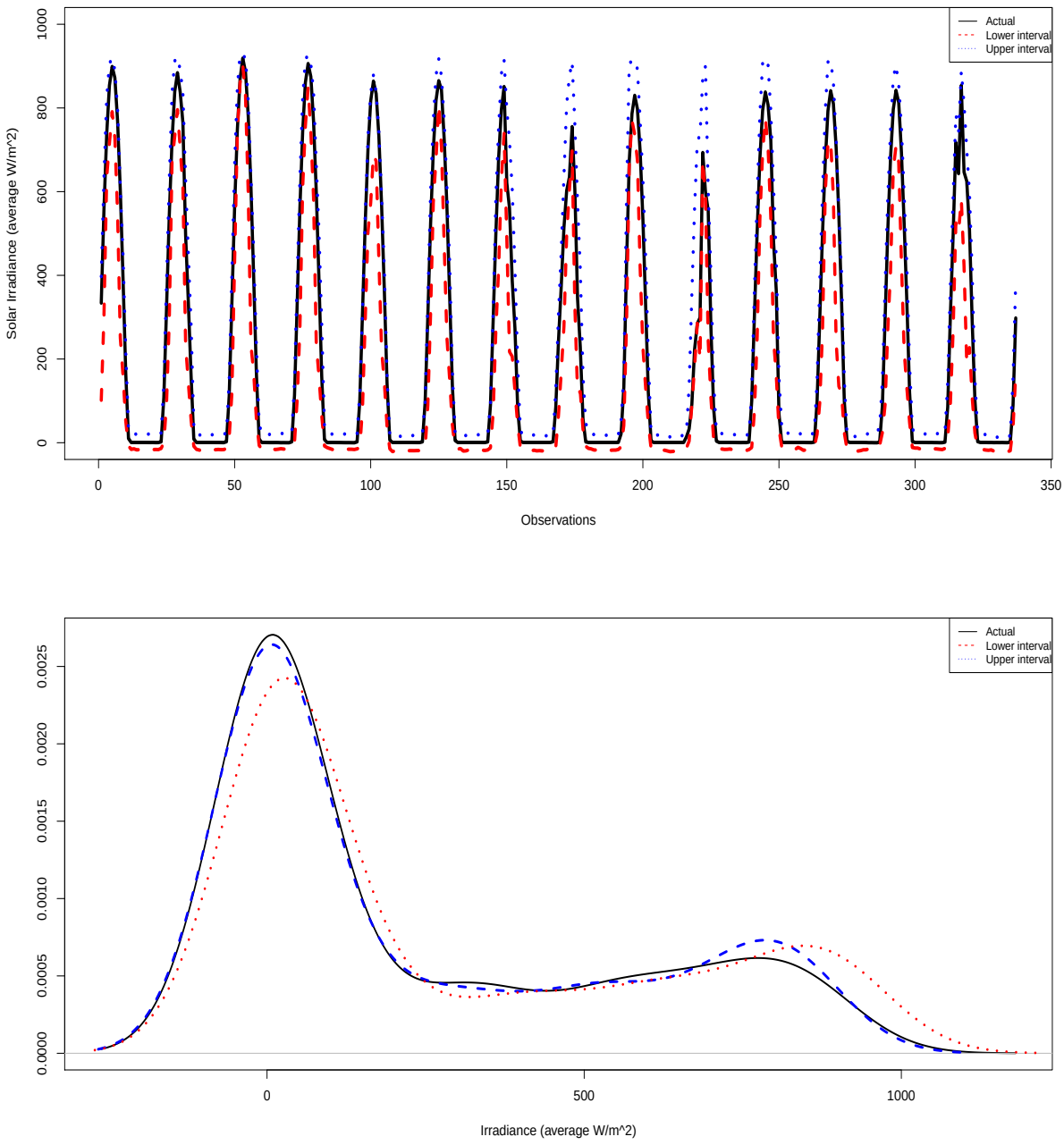

Figure 1: Top panel: Solar irradiance superimposed with the 1% and the 99% quantile forecasts. Bottom panel: Density plots for Solar irradiance superimposed with the 1% and the 99% quantile forecasts.

Appendix B

```
#########################################################################
# The following packages in R were used:                                ####
# forecast, ggplot2, qgam, mgcv, tseries, e1071, glmnet, hierNet, ####
#########################################################################

#########################################################################
### time series, qqnorn, density and box plot for solar irradiance ####
#########################################################################
win.graph()
A<-ts(Irrad)
par(mfrow=c(2,2))
plot(A,xlab="Observation number",ylab="Irradiance (average W/sgre m)"
                                    ,main="(a) Plot of irradiance")
plot(density(A),xlab="Irradiance
     ",main="(b) Density plot")
qqnorm(A,main="(c) Normal QQ plot")
qqline(A)
boxplot(A,main="(d) Box plot",varwidth=TRUE,horizontal= TRUE)
#########################################################################
## calculating summary statistics, skewness and kurtosis of solar ######
# irradiance                                                          ######
#########################################################################
summary(A)
```

```
library(e1071)
skewness(A)
kurtosis(A)
########################creating new variables ############################
#####nonlinear trend#################################################################
#####################################################################################
z = (smooth.spline(time(Irrad), Irrad))
z
lines(smooth.spline(time(Irrad), Irrad, spar=0.014408),col="red",lwd=3)
dpdfits = fitted((smooth.spline(time(Irrad), Irrad, spar=0.014408)))
plot(dpdfits)
write.table(dpdfits,"~/nolfits.txt",sep="\t")


#####################################################################################
# R code for variable selection using lasso and lasso via hierarchical ##
#interaction (using r-package glmnet and hierNet)                      ##
#####################################################################################
################ Variable Selection   ###################################
#####################################################################################
library(glmnet)
attach(site1analyticdata)
y<-site1analyticdata$Irrad
x<-data.matrix(site1analyticdata[3:12])
print(x)
```

```
LAsso<-glmnet(y=y,x=x,family="gaussian")
plot(LAsso)
coef(LAsso)
cv.lasso<-cv.glmnet(y=y,x=x,family="gaussian")
cv.lasso
plot(cv.lasso)
coef(cv.lasso,s="lambda.min")
predict.1<-predict(cv.lasso,newx=x)
predict.1
write.table(predict.1,"~/gam1.forecast2.txt",sep="\t")
m.lasso<-mean((y-predict.1)^2)
m.lasso
#############################################################################
##variable selection  via hierarchical interaction ######################
#############################################################################
library(hierNet)
attach(GAM_SolarDatae)
y<-GAM_SolarDatae$Irrad
x<-data.matrix(GAM_SolarDatae[3:12])
x
fit5<-hierNet(x,y,lam=50,strong=TRUE)
print(fit5)
# a typical analysis including cross-validation
fit6=hierNet.path(x,y, strong=TRUE)
fitcv=hierNet.cv(fit6,x,y)
```

```
win.graph()
print(fitcv)
plot(fitcv)
lamhat=fitcv$lamhat.1se
lamhat
fit1=hierNet(x,y,lam=lamhat, strong=TRUE)
print(fit1)
yhat=predict.hierNet(fit1,x)
yhat
write.table(yhat,"~/gam1.forecast2.txt",sep="\t")
plot(yhat)

#############################################################################
### R code for modeling GAM without interactions (using r-package mgcv)##
#############################################################################
#### The GAM model without interactions ###################################
#############################################################################
attach(site1analyticdata)
head(site1analyticdata)
win.graph()
idx_data_test<- 5362:nrow(site1analyticdata)
idx_data_test
data_train<- site1analyticdata[-idx_data_test, ]
data_train
data_test<- site1analyticdata[idx_data_test, ]
```

```
data_test
library(mgcv)
gam.fit1<- gam(Irrad~ s(RN)+s(WS)+s(WindDir)+s(AirTC)+s(RH)+s(BP)+
                 s(hour)+s(dif1Irad)+s(dif2Irad)
                 ,data= data_train)
##########################################################################
#### Ploting the model ###################################################
##########################################################################
win.graph()
par(mfrow=c(3,3))
plot(gam.fit1,shade=T)
par(mfrow=c(2,2))
gam.check(gam.fit1)
##########################################################################
############# Model Selection ############################################
##########################################################################
summary.gam(gam.fit1)
BIC(gam.fit1)
AIC(gam.fit1)
##########################################################################
####### Predicting Solar Irradiance#######################################
##########################################################################
gam.forecast<- predict.gam(gam.fit1, newdata= data_test)
gam.forecast
write.table(gam.forecast,"~/gam1.forecast.txt",sep="\t")
```

```
#########################################################################
############# Residuals #################################################
#########################################################################
win.graph()
library(forecast)
library(tseries)
r2<-residuals(gam.fit1)
r2
plot(r2)
tsdisplay(r2,main = "Residuals" )
Box.test(r2,lag=36,fitdf=2,type="Ljung")
#########################################################################
####### Correcting the autocorrelation ##################################
#########################################################################
fit4<-Arima(residuals.gam(gam.fit1),order=c(2,0,3)
         ,seasonal=list(order=c(1,0,1),period=24))
fit4
summary(fit4)
RES<-residuals(fit4)
#win.graph()
tsdisplay(RES,main="Residuals")
Box.test(RES,lag=20,fitdf=2,type="Ljung")
#########################################################################
############ Forecast Accuracy ##########################################
#########################################################################
```

```
library(forecast)
library(tseries)
attach(analyFGAM)
accuracy(GAMF,Irrad)
#########################################################################
########### Time Series Plot ############################################
#########################################################################
library(forecast)
library(tseries)
attach(analyFGAM)
win.graph()
z=ts(Irrad)
plot(z,col="blue",lwd=3,ylab="Irradiance (average W/m^2)",
     xlab="Observations",ylim=c(0,1000))
b=ts(GAMF)
lines(b,col="red",lty=2,lwd=3,ylim=c(0,1000))
legend("topright",legend=c("Actual","Forecast"),col=c("blue","red")
                                              ,lty=1:2,cex=0.8)
#########################################################################
############ Density plot  ##############################################
#########################################################################
library(forecast)
library(tseries)
attach(analyFGAM)
plot(density(z),col="red",xlab="Irradiance (average W/m^2)"
```

```
                ,ylab="",lwd=3,main="Density(Actual vs Forecast)")
lines(density(b),col="blue",lwd=3,lty=2)
legend("topright",legend=c("Actual","Forecast"),col=c("red","blue")
                                              ,lty=1:2,cex=0.8)
##############################################################################
####### Forecast combination r-codes using r-package opera  ##############
##############################################################################
library(opera)
y<-Irrad
x<-cbind(GAMFI,FQGAMI)
oracle.convex<-oracle(Y= y,experts = x,loss.type ="pinball"
                      ,model="convex")
win.graph()
plot(oracle.convex)
summary(oracle.convex)
mix1<-0.18*GAMFI + 0.82*FQGAMI
mix1
##############################################################################
## Pinball loss function using r-package gefcom2017  ######################
##############################################################################
library(gefcom2017)
pinballloss<-function(tau, y, q)
  pl-df <- data.frame(tau = tau,
                      y = y,
                      q = q)
```

```
pl-df <-pl-df mutate(L = ifelse(y>=q,tau/100*(y-q),
                                (1-tau/100)*(q-y)))
return(pl-df)
tau= 50
y= Act
q= mix1
z = pinball-loss(tau, y, q) z
write.table(z,"/pinballfplLassoI.txt",sep="")
qloss =L
a=ts(qloss)
plot(a)
mean(qloss)
```

References

M. Abuella and B. Chowdhury. Solar power probabilistic forecasting by using multiple linear regression analysis. In *SoutheastCon 2015*, pages 1–5. IEEE, 2015.

R. Adhikari and R.K. Agrawal. *An introductory study on time series modeling and forecasting*. arXiv preprint arXiv, 2013.

H. Akaike. Maximum likelihood identification of gaussian autoregressive moving average models. *Biometrika*, 60(2):255–265, 1973.

S. Alessandrini, D. MonacheL, S. Sperati, and G. Cervone. An analog ensemble for short-term probabilistic solar power forecast. *Applied Energy*, 157:95–110, 2015.

A.W. Aloysius, D. Yang, and W.F. Walsh. Day-ahead solar irradiance forecasting in a tropical environment. *Journal of Solar Energy Engineering*, 137(5):05–1009, 2015.

P. Bacher, H. Madsen, and H.A Nielsen. Online short-term solar power forecasting. *Solar Energy*, 83:1772–1783, 2009.

J.M. Bates and C.W.J. Granger. The combination of forecasts. *Journal of the Operational Research Society*, 20(4):451–468, 1969.

J. Bien, J. Taylor, and R. Tibshirani. A lasso for hierarchical interactions. *Annals of statistica*, 41(3):1111–1141, 2013.

G. Bontempi and T.S. Ben. Statistical foundations of machine learning. *Universite Libre de Bruxelles*, 2011.

M. Bouzerdoum, A. Mellit, and A.M. Pavan. A hybrid model (sarima–svm) for short-term power forecasting of a small-scale grid-connected photovoltaic plant. *Solar Energy*, 98:226–235, 2013.

J.B. Bremnes. Probabilistic wind power forecasts using local quantile regression. *Wind Energy*, 7(1):47–54, 2004.

M.J Brooks, C.S Du, W.L. Van Niekerk, P. Gauché, C. Leonard, M.J. Mouzouris, R. Meyer, N. Van der Westhuizen, E.E Van Dyk, and F.J.Vorster. Sauran: A new resource for solar radiometric data in Southern Africa. *Journal of Energy in Southern Africa*, 26(1):2–10, 2015.

G. Casella and R.L. Berger. *Statistical inference*, volume 2. Pacific Grove, CA:Duxbury, 2002.

D. Chikobvu and C. Sigauke. Regression-sarima modelling of daily peak electricity demand in south africa. *Journal of Energy in Southern Africa*, 23(3):23–30, 2012.

C.F.M. Coimbra and R. Marquez. Proposed metric for evaluation of solar forecasting models. *Journal of solar energy engineering*, 135(1):011–016, 2013.

P. Craven and G. Wahba. Smoothing noisy data with spline functions. *Numerische mathematik*, 31(4):377–403, 1978.

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

J. Dekker, M. Nthontho, S. Chowdhury, and S.P Chowdhury. Investigating the effects of solar modelling using different solar irradiation data sets and sources within South Africa. *Solar Energy*, 86:2354–2365, 2012.

M. Devaine, P. Gaillard, Y. Goude, and G. Stoltz. Forecasting the electricity consumption by aggregating specialized experts: A review of sequential aggregation of specialized experts, with an application to slovakian and french country-wide one-day-ahead (half-) hourly predictions. *Machine Learning*, 90:231–260, 2012.

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.

C.F. Dormann, J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carré, J.R. Marquéz, B. Gruber, B. Lafourcade, P.J. Leitão, et al. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1):27–46, 2013.

A.T. Eseye, J. Zhang, and D. Zheng. Short-term photovoltaic solar power forecasting using a hybrid wavelet-pso-svm model based on scada and meteorological information. *Renewable Energy*, 118:357–367, 2018.

F.J Fabozzi, S.M. Focardi, S.T. Rachev, and B.G. Arshanapalli. *The Basics of Financial Econometrics: Tools, Concenpts, and Asset Management Application.* New Jersey: John Wiley and Sons, 2014.

G. Faranak, P. Pierre, and H.B. Gooi. Very short-term nonparametric probabilistic forecasting of renewable energy generation - with application to solar energy. *IEEE Transactions on Power Systems*, 31(5):3850–3863, 2016.

P. Gaillard and Y. Goude. Forecasting electricity consumption by aggregating experts; how to design a good set of experts. pages 95–115, 2015.

A. Grantham, R.Y. Gel, and J.W. Boland. *Nonparametric short-term probabilistic forecasting for solar radiation. Solar energy*, 133:465–475, 2016.

W. Hardle, Z. Hlávka, and S. Klinke. *XploRe: Application Guide.* Springer Berlin Heidelberg, 2000.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*, volume 1. Springer Series in Statistics springer, Berlin, 2001.

Trevor Hastie and Robert Tibshirani. *Generalized additive models.* Wiley Online Library, 1990.

B. Hodge, D. Lew, M. Milligan, H. Holttinen, S. Sillanpää, E. Gómez-Lázaro, R. Scharff, L. Söder, X.G Larséand G. Giebel, et al. Wind power forecasting error distributions: An international comparison. In *11th Annual International Workshop on Large-Scale Integration of Wind Power into Power Systems as well as on Transmission Networks for Offshore Wind Power Plants Conference*, 2012.

T. Hoshino. Quantile regression estimation of partially linear additive models. *Journal of Nonparametric Statistics*, 26(3):509–536, 2014.

Y. Hu, K. Zho, and H. Lian. Bayesian quantile regression for partially linear additive models. *Statistics and Computing*, 25(3):651–668, 2015.

J. Huang and M. Perry. A semi-empirical approach using gradient boosting and k-nearest neighbors regression for gefcom2014 probabilistic solar power forecasting. *International Journal of Forecasting*, 32(3):1081–1086, 2016.

R.J. Hyndman and G. Athanasopoulos. Forecasting: principles and practice. *The book is freely available as an online book at www.otexts. org/fpp.Alternatively, a print version is available: ISBN*, 987507109, 2014.

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*, volume 112. New York: springer, 2013.

R. Juban, H. Ohlsson, M. Maasoumya, L. Poirier, and J.Z. Kolter. A multiple quantile regression approach to the wind, solar, and price tracks of gefcom2014. 32:1094–1102, 2016.

J.Zhang, B.M. Hodge, and A. Florita. Investigating the correlation between wind and solar power forecast errors in western interconnection:preprint. Technical report, 2013a.

J.Zhang, B.M. Hodge, A. Florita, S. Lu, H. Hamann, and V. Banunaraynan. Metrics for evaluating the accuracy of solar power forecasting (presentation). Technical report, 2013b.

E.G. Kardakos, M.C. Alexiadis, S.I. Vagropoulos, C.K. Simoglou, P.N. Biskas, and A.G.Bakirtzis. Application of time series and artificial neu-

ral network models in short-term forecasting of pv power generation. In *Power Engineering Conference (UPEC), 2013 48th International Universities*, pages 1–6. IEEE, 2013.

A.E. Hoerland R.W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

R. Koenker. *Quantile regression*. Number 38. New York: Cambridge university press, 2005.

A.C. Kruger and S.S. Sekele. Trends in extreme temperature indices in South Africa: 1962-2009. *International Journal of Climatology*, 33:661–676, 2013.

K. Larsen. Gam: The predictive modelling silver bullet. *Multithreaded. Stitch Fix*, 30:1–27, 2015.

M. Lim and T. Hastie. Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics*, 24(3): 627–654, 2015.

K. Maciejowska, J. Nowotarski, and R. Weron. Probabilistic forecasting of electricity spot price using factor quantile regression averaging. *International Journal of forecasting*, 32(3):957–965, 2016.

D. Marasinghe. *Quantile regression for climate data*. Master's thesis, Clemson University, 2014.

A.A. Mohammed, W. Yaqub, and Z. Aung. Probabilistic forecasting of solar power:an ensemble learning approach. In *Intelligent Decision Technologies*, pages 449–458. Springer, 2015.

H.A. Nielsen, H. Madsen, and T.S. Nielsen. Using quantile regression to extend an existing wind power forecasting system with probabilistic forecasts. *Wind Energy*, 9(1-2):95–108, 2006.

J. Nowotarski and R. Weron. Computing electricity spot price prediction intervals using quantile regression and forecast averaging. *Computational Statistics*, 30(3):791–803, 2015.

J.X. Pan and K.T. Fang. *Growth curve models and statistical diagnostics*. Springer Science & Business Media, 2012.

G. Pierre, G. Yannig, and N. Raphaël. Additive models and robust aggregation for gefcom2014 probabilistic electric load and electricity price forecasting. *International Journal of Forecasting*, 32(3):1038–1050, 2016.

S. Portnoy and R. Koenker. The Gaussian hare and the Laplacian tortoise: Computability of squared- error versus absolute-error estimators. *Statistical Science*, 12(4):279–296, 1997.

W.G. Le Roux. Optimum tilt and azimuth angles for fixed solar collectors in South Africa using measured data. *Renewable Energy*, 96:603–612, 2016.

G. Seni and J.F. Elder. Ensemble methods in data mining: improving accuracy through combining predictions. *Synthesis Lectures on Data Mining and Knowledge Discovery*, 2(1):1–126, 2010.

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.

E. Sjoborg. *Modelling and Forecasting Electricity Load in Secondary Substations*. Master's thesis, Lund University, 2017.

A. Sozen, A. Mirzapour, and M.T. Çakir. Selection of the best location for solar plants in turkey. *Journal of Energy in Southern Africa*, 26(4):52–63, 2015.

X. Sun, Z. Wang, and J. Hu. Prediction interval construction for byproduct gas flow using optimizing twin extreme learning machine. *Mathematical Problems in Engineering*, 2017:1–12, 2017.

H. Tao, P. Pierre, F. Shu, and Z. Hamidreza. Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond. *International Journal of Forecasting*, 32:896–913, 2016.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

G.A. Warner. Solar energy in South Africa: Challenges and opportunities, 2014. URL http://www.ee.co.za/article/solar-energy-south-africa-challenges-opportunities.html. [Online; accessed 5-18-2017].

S.N. Wood. *Generalized Additive Models: An Introduction with R*. New York: Taylor and Francis, 2 edition, 2017.

D. Yang, Z. Ye, H.I. Lim Li, and Z. Dong. Very short term irradiance forecasting using the lasso. *Solar Energy*, 114:314–326, 2015.

E. Zhandire. Predicting clear-sky global horizontal irradiance at eight locations in south africa using four models. *Journal of Energy in Southern Africa*, 28(4):77–86, 2017.

J. Zhang, A. Florita, B. Hodge, S. Lu, H.F. Hamann, V. Banunarayanan, and A.M. Brockway. A suite of metrics for assessing the performance of solar power forecasting. *Solar Energy*, 111:157–175, 2015.