Statistical Modelling and Estimation of Solar Radiation

by

Mphiliseni Bongani Nzuza

Submitted in fulfilment of the academic requirements for the degree of
Master of Science
in the Discipline of Statistics
School of Mathematics, Statistics and Computer Science
THE UNIVERSITY OF KWAZULU-NATAL
Durban

UNIVERSITY OF
KWAZULU-NATAL

March 2014

Preface

The experimental work described in this thesis was carried out in the School of Mathematics, Statistics & Computer Science, University of KwaZulu-Natal, Durban, from July 2011 to December 2013, under the supervision of Doctor E. Ranganai and Professor G. Matthews.

These studies represent original work by the author and have not otherwise been submitted in any form for any degree or diploma to any tertiary institution. Where use has been made of the work of others it is duly acknowledged in the text.

Signed:
M Nzuza (candidate)

Signed:
Doctor E Ranganai (supervisor)

Signed:
Professor G Matthews (co-supervisor).

Declaration - Plagiarism

I, Mphiliseni Bongani Nzuza, declare that

1. The research reported in this thesis, except where otherwise indicated, is my original research.

2. This thesis has not been submitted for any degree or examination at any other university.

3. This thesis does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.

4. This thesis does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:

   a. Their words have been re-written but the general information attributed to them has been referenced.

   b. Where their exact words have been used, then their writing has been placed in italics and inside quotation marks, and referenced.

5. This thesis does not contain text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the thesis and in the References sections.

Signed:

Acknowledgements

Special thanks go to my supervisor, Doctor Edmore Ranganai, for his careful dedication to my work and to my co-supervisor, Professor Glenda Matthews, for introducing me to the field of solar energy resource studies, helping me with relevant study material and thesis writing. I am duly indebted to the College of Science and Agriculture as well as the National Research Foundation for their financial support. I am also grateful to Michael Brooks for providing data and relevant information that have contributed to the success of this study. Lastly, I would like to thank everyone else in the School of Statistics for their encouragement and support.

Abstract

Solar radiation is a primary driving force behind a number of solar energy applications such as photovoltaic systems for electricity generation amongst others. Hence, the accurate modelling and prediction of the solar flux incident at a particular location, is essential for the design and performance prediction of solar energy conversion systems. In this regard, literature shows that time series models such as the Box-Jenkins Seasonal/Non-seasonal Autoregressive Integrated Moving Average (S/ARIMA) stochastic models have considerable efficacy to describe, monitor and forecast solar radiation data series at various sites on the earth's surface (see e.g. Reikard, 2009). This success is attributable to their ability to capture the stochastic component of the irradiance series due to the effects of the ever-changing atmospheric conditions. On the other hand at the top of the atmosphere, there are no such conditions and deterministic models which have been used successfully to model extra-terrestrial solar radiation. One such modelling procedure is the use of a sinusoidal predictor at determined harmonic (Fourier) frequencies to capture the inherent periodicities (seasonalities) due to the diurnal cycle. We combine this deterministic model component and SARIMA models to construct harmonically coupled SARIMA (HCSARIMA) models to model the resulting mixture of stochastic and deterministic components of solar radiation recorded at the earth's surface. A comparative study of these two classes of models is undertaken for the horizontal global solar irradiance incident on the solar panels at UKZN Howard College (UKZN HC), located at 29.9º South, 30.98º East with elevation, 151.3m. The results indicated that both SARIMA and HCSARIMA models are good in describing the underlying data generating processes for all data series with respect to different diagnostics. In terms of the predictive ability, the HCSARIMA models generally had a competitive edge over the SARIMA models in most cases. Also, a tentative study of long range dependence (long memory) shows this phenomenon to be inherent in high frequency data series. Therefore autoregressive fractionally integrated moving average (ARFIMA) models are recommended for further studies on high frequency irradiance.

Contents

List of Figures

List of Tables

List of Abbreviations

AIC	Akaike’s Information Criteria
ANN	Artificial Neural Networks
AR	Autoregressive
ARCH	Autoregressive Conditional Heteroscedasticity
ARFIMA	Autoregressive Fractionally Integrated Moving Average
BIC	Bayesian Information Criteria
CARDS	Coupled Autoregressive and Dynamical
CI	Confidence Interval
Corr	Correlation
Cov	Covariance
DHI	Diffuse Horizontal Irradiance
DNI	Direct Normal Irradiance
Eq.	Equation
ES	Exponential Smoothing
GARCH	Generalized Autoregressive Conditional Heteroscedasticity
GHI	Global Horizontal Irradiance
GRADRAD	Greater Durban Radiometric Network
HCSARIMA	Harmonically Coupled Seasonal Autoregressive Integrated Moving Average
IID	Independent and Identically Distributed
LRD	Long Range Dependence
MA	Moving Average
MAPE	Mean Absolute Percentage Error
MBE	Mean Bias Error
MPE	Mean Percentage Error
MVN	Multivariate Normal
NDD	Normalized Discrete Difference
NID	Normally and Independently Distributed
NIP	Normal Incident Pyrheliometer
PSP	Precision Spectral Pyranometer
PV	Photovoltaic
RMSE	Root Mean Square Error
S/ARIMA	Seasonal Autoregressive Integrated Moving Average
SARFIMA	Seasonal Autoregressive Fractionally Integrated Moving Average
SBC	Coupled Autoregressive and Dynamical System Schwarz’s Bayesian Criterion
SES	Simple Exponential Smoothing
SRA	Sample Residual Autocorrelation
SRPA	Sample Residual Partial Autocorrelation
TES	Two-Directional Exponential Smoothing
Var	Variance

Chapter 1
Introduction

Preliminaries

In this preliminary chapter, a short motivation of the study as well as little background on solar energy studies is given. A detailed explanation of the aims and objectives of the study is also provided, as well as a brief introduction to some of the important concepts in the solar energy studies.

1.1 Motivation

The increasing consumption of solar power as a source of electricity creates a greater need in assessing and predicting solar resource over various time horizons, depending on the requirements. Among many services, short-term energy forecast information is essentially required for operational planning, switching sources, programming back-up, short-term power purchases, planning for reserve usage and peak load matching. The growing number of solar systems installations worldwide is an indication that the accurate assessment of solar resource is essential to facilitate the design of solar electric grids. Therefore, solar irradiance quantification studies are of great significance for the optimal operation and power prediction of grid connected photovoltaic (PV) plants. However, this presents a challenge which is very complex to handle due to the random and nonlinear characteristics of solar irradiance under changeable weather conditions. Such uncertainties associated with the variations of solar flux incident on the solar panels leave much to be desired. Thus, the uncertainty quantification of the stochastic (random) variations of solar irradiance might be one essential step, as an efficient use of solar resource requires reliable information related to its availability.

Short term solar irradiance forecasting (up to a few minutes or hours or days) has significant aids in solar energy system sizing and optimization and is therefore critical for solar system developers. Accurate forecast information improves the efficiency of the solar systems outputs. The importance of solar resource forecasting can be witnessed in energy storage management of stand-alone photo-voltaic (PV) or wind energy systems, control systems in buildings, control of solar thermal power plants and the management of electricity grids with high penetration rates from renewable sources. Both physical and statistical models have been used to assess solar radiation at the earth’s surface (see e.g. Badescu et al., 2008). However, the need for reliable predictive methods for solar systems power output arises, for instance, in operational planning procedures related to future energy availability, demand etc. The findings of research studies in the field of solar energy could be useful in solar systems development and power management.

The general class of models have been adapted or coupled with other model forms to deal with some data phenomena deviating from the classical assumptions (“norms”). For example, the time series data with deterministic seasonal patterns and autocorrelated errors can be modelled by a deterministic regressor and the residuals by a S/ARIMA model. In the case of long range dependence inherent in the series, the Autoregressive Fractionally Integrated Moving Average (ARFIMA) models have been with effect (see e.g. Granger and Joyeux, 1980). Therefore, in this thesis we make use of S/ARIMA related models.

1.2 Background Studies

Some attempts have been made previously to quantify the uncertainties associated with the variations of solar irradiance incident on the ground. The earliest studies in the field of solar energy were conducted by Liu and Jordan (1960). These researchers established the relationship between daily diffuse and global irradiance components on clear days on a horizontal surface, with the measurements from 98 sites in the US and Canada. In an attempt to assess global solar irradiance, various classes of models such as regressions in logs, seasonal autoregressive integrated moving average (S/ARIMA), transfer functions, neural networks (see e.g. Alam et al., 2006; Tymvios et al., 2005), and hybrid models have been employed amongst others. Some research studies have been carried out examining global solar irradiance at various resolutions ranging from about 5 minutes to as long as a day (see e.g. Craggs et al., 1999); Reikard, 2009).

The success of S/ARIMA is attributable mainly to its ability to capture the cycles more effectively than other methods. For example, this was evident in the findings of the study by Craggs et al., (1999) to test the efficacy of S/ARIMA models in evaluating the 60-minutely and 10-minutely averaged global horizontal irradiance relating to 13 and 15 day periods in two winters and two summers. In the aforementioned study, the S/ARIMA models were used for short-term prediction of irradiance at the northerly location in the city centre of Newcastle upon Tyne, UK at latitude 54859'N, longitude 1837'W and altitude 44m. In this study, a univariate stochastic modelling using S/ARIMA models was carried out for horizontal and south facing vertical solar irradiance. The results showed that these models provide a good fit for the 10-minutely averaged horizontal and vertical irradiance. However, the use of the 60-minutely averaged data in these models gave a substantial reduction in the fit. In another study by Reikard (2009), in an attempt to estimate the global horizontal solar irradiance, the data series were examined at resolutions ranging from about 5 min to 60 min. The results of this study indicated that neural networks or hybrid models in a few cases can improve at very high resolutions on the order of 5 min while the success of the S/ARIMA models was attributable mainly to its ability to capture the diurnal cycle more effectively than other methods. For variance stabilizing purposes, the models were fitted to the log transform of the original series. Overall, both studies indicated that the best results were achieved from S/ARIMA in logs.

Furthermore, one of the recent studies has been based on measurements of global solar radiation from the National University of Colombia in Bogota (74º 4' West, 4º 35' North, 2580 m altitude) for the period from 2003 to 2009 (see Perdomo et al., 2010). In this study, a time series statistical modelling has been performed in an attempt to predict the accumulated mean daily global solar radiation at the solar station of National University of Colombia in Bogota. The stationarized version of the data series was examined and the ARIMA (1,0,0) was employed as a best fit, with the error term distributed as a standard normal variable (i.e. white noise). Also deterministic models have been used to model and predict solar irradiance. One such approach is the application of sinusoidal prediction techniques (see e.g. Huang et al., 2011). In this thesis, we couple these predictors with S/ARIMA models to form harmonically coupled S/ARIMA (HCSARIMA) models.

1.3 Aims and Objectives

The aim of this study is detailed as follows:

The electrical output from a photovoltaic (PV) panel in the horizontal plane on the earth's surface is influenced by the variable daily meteorological conditions and hence uncertainties due to random variations of solar resource. Therefore, reliable forecasts are critical to solar system developers because of the future uncertainties about the performance of a system. In particular, the aim is to clarify the exact nature of solar irradiance falling on the radiometric ground station of UKZN HC so that the forecasting may be performed by a specified stochastic model.

Another challenge faced is the estimation of missing values in the measured solar data caused by various phenomena such as equipment malfunction and interruptive maintenance among others. This is inherent in many datasets containing gaps e.g., data recorded at UKZN HC Solar Meteorological Station, which we make use of in this study. Apart from that, measuring instrumentation can be anticipated to fail from time to time and therefore be faulty to give infeasible values (with high error margin) or no values at all. For this reason, estimation models for solar radiation are required for efficiently monitoring solar system. In this thesis we study global horizontal irradiance (GHI) although its components, namely, direct normal irradiance (DNI) and diffuse horizontal irradiance (DHI) are also recorded.

To our knowledge, the aforementioned classes of models, namely; SARIMA, HCSARIMA and ARFIMA have never been used to assess solar resource incident at the solar station (UKZN HC) under investigation, nor have any studies of this nature been carried out at this station. This is therefore one of the contributions of this study. The second contribution of the study is to be able to predict the irradiance pattern for the site with some degree of accuracy. The designer of solar energy collection systems may be interested in knowing how much solar energy he anticipates to fall on a collector over a certain period of time such as a day or two. If storage is included in a system design, the designer also needs to know the variation of solar irradiance over time for system design and optimization purposes, in which case the predictive models can be searched and formulated to assist the designer achieve that. Hence this will enable us to tell the designer the next irradiance pattern to expect within a couple of periods at UKZN HC. A situation of this kind has prompted the development of efficient models to provide reliable irradiance predictions in an attempt to estimate the missing values for solar stations where the equipment malfunction is experienced from time to time. Therefore, this could be the primary interest, i.e. the interpolation of missing values prior to data modelling.

Finally, we concentrate on searching for the models which best explain the underlying data generating processes for irradiance time series data obtained from the station of UKZN HC. This is also intended to improve on some previously used methods for estimating irradiance time series data with certain disadvantages (see e.g. Wang et al., 2012). This researcher made use of extra-terrestrial irradiance model to estimate global solar irradiance at the ground level. Such approach has clear disadvantage in that it only takes into consideration geographic quantities thereby providing estimates with a high margin of error even on clear sky days, overestimating and/or underestimating. Apparently, such models represent no underlying stochastic process of the series because it is not developed from the sample. Using different datasets, we show that there are better methods of modelling solar irradiance patterns on the ground. Time series models are capable of capturing the stochastic (random) component infused in an irradiance time series data of all types of weather, providing better estimation. We also assess whether employing Harmonically Coupled SARIMA (HCSARIMA) models yields better results. At a tentative level, we carry out a study of long range dependence in the irradiance series and point out areas of further research.

Therefore the main objectives of this study are summarized as follows:

- Generally searching for the most accurate underlying data generating processes that could be used for the generation of series values with a higher degree of accuracy and to replace some of the previously used less effective methods.
- Modelling solar irradiance using advanced time series analysis techniques e.g. Box-Jenkins SARIMA.
- Combining sinusoidal component inherent in the series and SARIMA models to form a new class of models namely Harmonically Coupled Seasonal Autoregressive Integrated Moving Average (HCSARIMA) processes, which are also used to model the same solar irradiance datasets.
- Comparing the performance of SARIMA and HCSARIMA models in terms of their competencies to forecast for solar irradiance series, on the basis of forecast error (accuracy) measures.
- Preliminarily showcasing the ability of the ARFIMA process to model the high frequency time series data.

1.4 Fundamental Concepts and KZN Solar Distribution

For the purpose of solar power supply, the most significant measures are the intensity and energy delivered, hence one measure at a point in time and the other over a period of time. The *rate* at which the solar energy reaches a unit area at the earth's surface is called the "*solar irradiance*" or "*insolation*". It is the intensity of solar radiation hitting a surface, which is the sum of the contributions of all the wavelengths within the spectrum. The units of measurement for irradiance are watts per square meter (W/m^2). In simpler terms, solar irradiance can be defined as an *instantaneous* measure of rate and is variable over time. The maximum solar irradiance value is used in system design to determine the peak rate of energy input into the system. The *solar irradiation or radiation* is simply the *integration* or *summation* of *solar irradiance* over a time period. For instance, let us consider the irradiance incident on a unit area over a finite time interval $\Delta t = t_2 - t_1$, then the respective energy realized on this unit area can be defined for irradiation as follows:

$$H = \int_{\Delta t} I(t) dt \qquad (1.1)$$

where $I(t)$ is the solar irradiance value at time instant t. The common measurement units of irradiation H are J/m^2 (Joule per square meter) or Wh/m^2(Watt-hours per square meter). The momentary total irradiance incident on a solar collector is generally referred to as power, measured in watts ($W = Js^{-1}$), i.e. the rate at which the work is done (see e.g. Watt, 1978).

The radiation intensity on the surface of the sun is approximately $6.33 \times 10^7 W/m^2$ and the intensity of the radiation leaving the sun is relatively constant. It is the amount of energy received at the top of the earth's atmosphere, measured at an average distance between the earth and sun on a surface oriented perpendicular to the sun. As it travels to the earth's surface, the radiation spreads out as the distance squared bringing about the reduction of the radiant energy falling on $1m^2$ of surface area to a constant I_0, called the *solar constant* (see e.g. Froehlich and Brusa, 1981; Iqbal, 1983), with the generally accepted value of $1367W/m^2$. A solar map of KwaZulu-Natal Global Horizontal Irradiation given below in Figure1.1, shows that Durban possesses a considerable solar resource of approximately $1637kWh/m^2$, annually. It is notable that we experience a higher concentration of solar flux as we move farther away from the coastal regions.

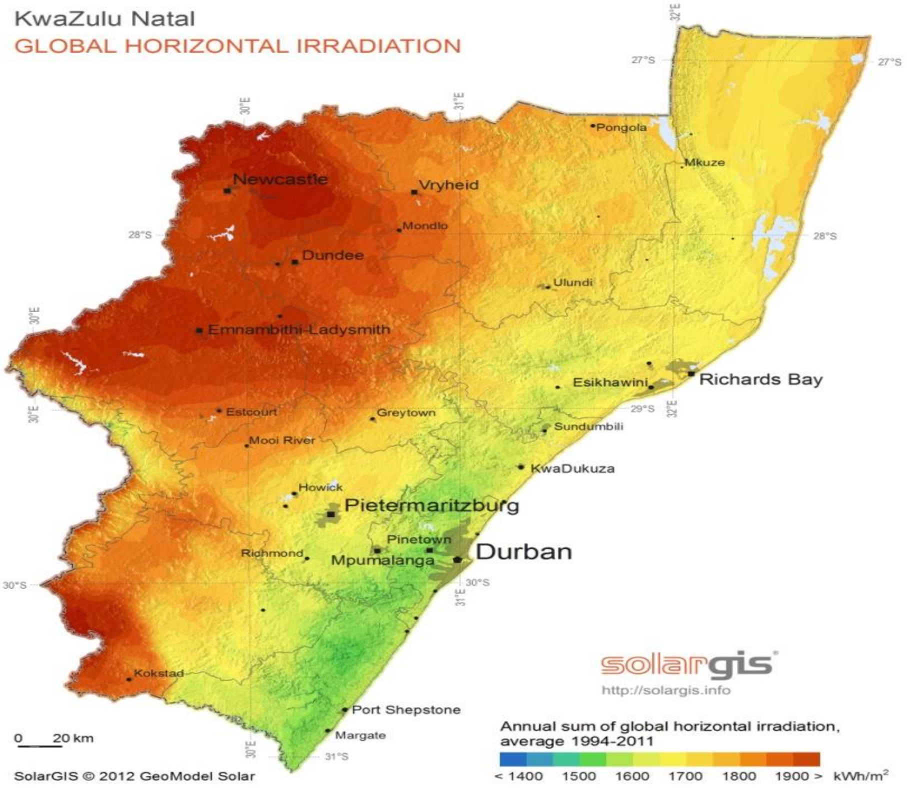

Figure 1.1: A solar map of KwaZulu-Natal Global Horizontal Irradiation. Source: www.kzngreengrowth.com.

1.5 Thesis structure

In this section an outline of the remaining chapters is given to summarize each chapter's content.

Chapter 2 introduces the background research studies in the field of solar energy. From this chapter, we gain an understanding of how the incoming energy from the sun is influenced by meteorological factors as it traverses the atmosphere to the ground. The physical models have been developed in an effort to estimate solar radiation received on the ground. In this chapter we also give some scientific time series models that can be used in addressing the challenges arising in design and sizing of solar power systems as well as power management of such systems. In Chapter 3 we discuss in detail the two main approaches to analysing time series data, namely, *time domain* and *frequency domain* techniques. The first approach (time domain) generally makes use of the general Box-Jenkins techniques in building a model. The latter approach (frequency domain) is appropriate when fluctuations of sinusoidal nature are inherent in the series. Spectral analysis of the series is then carried out to search for periodicities within the data. Chapter 4 gives a detailed discussion of the long memory (long range dependence) property inherent in high frequency time series data. This is characterized by autocorrelations that decay very slowly or fail to decay at an earlier lags. For this reason, a special class of models viz., Autoregressive Fractionally Integrated Moving Average (ARFIMA) models, has been proposed in an effort to deal with the long memory dependence. The ARFIMA process allows non-integer (fractional) values of the differencing parameter. In Chapter 5, various forecasting methods are discussed with respect to their application according to specific behaviours by time series data. Such forecasting techniques are moving average and simple exponential smoothing methods, double exponential smoothing, triple exponential smoothing, multiplicative and additive seasonal models. In Chapter 6, we discuss data availability, measurement techniques, the missing data problem and data modelling. Finally, in Chapter 7 we give a detailed conclusion on the research findings and also point out some areas for further research.

Chapter 2

Review of Literature on Solar Irradiance

2.1 Solar Irradiance Components

As solar irradiance traverses the atmosphere in the form of electromagnetic waves or sun's rays, some of it can be reflected, absorbed, scattered and transmitted by an intervening medium such as air molecules or clouds. This occurs in varying amounts depending on the wavelength. As a consequence, the solar input into the earth’s surface is reduced and falls on a solar panel in various forms. The complex interactions of solar irradiance with the earth's atmosphere result in the fundamental broadband components, namely, *beam or direct irradiance*, denoted by I_b, and *diffuse or scattered irradiance*, denoted by I_d, on which information is needed for solar energy conversion technologies. These sources add up to the total which is referred to as *global or total solar irradiance*, denoted by I_g. However, at the stage of data modelling, we denote the irradiance time series by Y_t.

On the surface of the earth, we perceive the beam or direct solar irradiance that comes directly from the sun and the diffuse or scattered solar irradiance that appears to come from various directions over the entire sky due to atmospheric scattering. Thus, the term "global" is associated with the fact that the solar irradiance on a horizontal surface is received from the entire 2π solid angle of the sky dome. Direct irradiance can also be reflected by the surrounding environment on to a solar device or panel. This is called *ground-reflected solar irradiance* (see Figure 2.1).

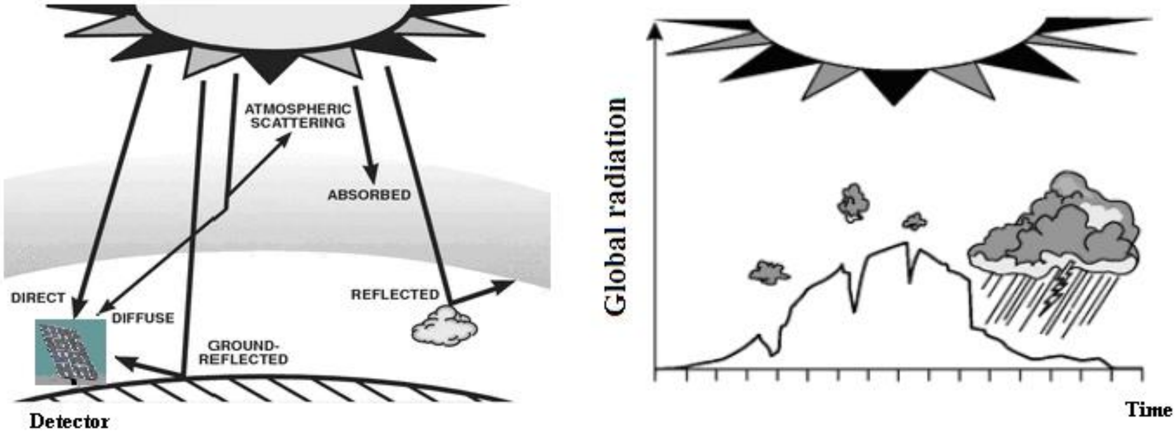

Figure 2.1: Radiation scattering and reduction, three types of radiation: direct, diffuse and ground reflected. Source: http:// www.newport.com/Introduction-to-Solar-Radiation.

It is also observable that some portion of energy is backscattered by the atmosphere and some reflected by the cloud cover before reaching the ground. Meanwhile, this allows us to conclude that the difference observed between global irradiance on a detector at ground level and its corresponding value outside the atmosphere is what has been absorbed, backscattered or reflected away. In the following section, a bit of basic physical modelling behind solar radiation is given.

2.2 Extra-terrestrial Solar Irradiance and Cosine Effect

The *extra-terrestrial solar irradiance* is an instructional concept often used in solar irradiance deterministic models. This is not affected by the atmospheric or weather conditions. Rather, it is determined by the earth's rotation and revolution. That is, outside the atmosphere, this intensity varies only due to the earth's orbit being slightly elliptical. It changes with the day of the year and the *maximum* irradiance occurs at the *perihelion* i.e. the earth closest to the sun (sometime in January) and the *minimum* at the aphelion (sometime in July). This variation is expressed in terms of the eccentricity correction factor ϵ_0 as follows:

$$\epsilon_0 = 1 + 0.033\cos\left(\text{RADIANS}\left(\frac{360N}{365}\right)\right), \tag{2.1}$$

where $N \in [1, 36]$ is the day number, starting from the 1st of January (see e.g. Badescu, 2008; Iqbal, 1983).

From Equation (2.1), the extra-terrestrial irradiance at a normal incidence is given by

$$I_0 = I_{sc}\epsilon_0 \tag{2.2}$$

The Lambert' Law (*Cosine Effect*): Since the sunlight is smoothly distributed over whole areas, a mere figure for intensity is never sufficient without knowledge of the orientation of the surface in question. Typically, the orientation of a surface is described by the zenith angle,θ_z, the angle between the sunbeam and the normal of the area. If I_0 is the extra-terrestrial solar irradiance (i.e. the irradiance initially available at the top of the atmosphere) falling on a horizontal surface, the intensity on an area where the sun is observed under the zenith angle θ_z, is given by

$$I_{0,h} = I_0 \cos(\theta_z)\,,\ 0^{\circ} \leq \theta_z \leq 90^{\circ}. \tag{2.3}$$

This means that if the surface is perpendicular to the sunbeam (normal to a central ray), i.e. $\theta_z = 0$, the solar irradiance falling on it will be I_0, the maximum possible solar irradiance. On the other hand, if the surface area is not perpendicular to the sunbeam, it is notable that a larger area may be required to catch the same flow as the cross section of the sunbeam. Equation (2.3) is generally referred to as Lambert's Law (see e.g. Baldocchi, 2012), named after Johann Heinrich Lambert, from his Photometria (1760). The cosine effect and/or Lambert's Law is diagrammatically described in Figure 2.2.

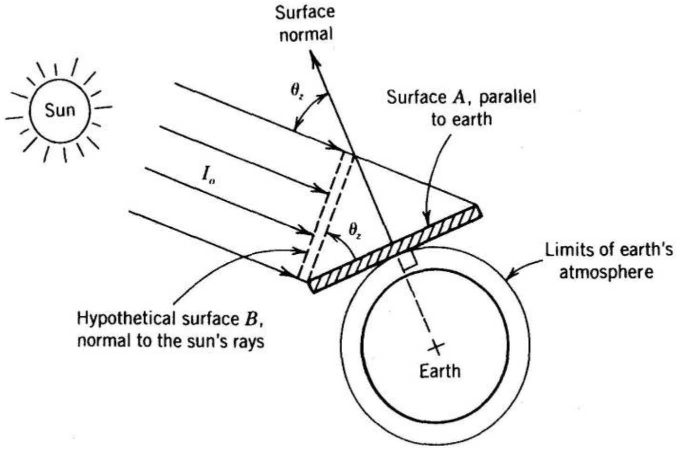
Surface normal
θ_z
Sun
Surface A, parallel to earth
I_o
θ_z
Limits of earth's atmosphere
Hypothetical surface B, normal to the sun's rays
Earth

Figure 2.2: The cosine effect as it relates to the concept of extra-terrestrial irradiance on a horizontal surface. *Source*: http://www.powerfromthesun.net/chapter2/Chapter2.htm.

The effect of geographical quantities: For a particular location, on a particular day in a year, the extra-terrestrial irradiance I_0 can be deterministically estimated as a function of basic geographic and astronomic quantities such as latitude (ϕ), declination (δ) and hour angle (ω) among others, (see e.g. Radosavljevic and Dordevic, 2001). The cosine of the solar zenith angle (θ_z) can be expressed in terms of the aforementioned quantities as follows,

$$\cos(\theta_z) = \sin\delta \sin\phi + \cos\phi \cos\omega \cos\delta. \tag{2.4}$$

Therefore, by substituting Equation (2.4) into Equation (2.3), the intensity of extra-terrestrial radiation on horizontal surface for particular day in a year can better be estimated by the following formula:

$$I_0 = I_{sc}\left[1 + 0.033\cos\left(\text{RADIANS}\left(\frac{360N}{365}\right)\right)\right](\sin\delta \sin\phi + \cos\phi \cos\omega \cos\delta). \tag{2.5}$$

Various aerosol factors such as clouds thickness and water vapour among others bring about reduction of solar energy as it traverses to the surface in the form of electromagnetic waves, the energy received on the ground in a less amounts than expected extra-terrestrial value. Therefore, the difference between extra-terrestrial irradiance and surface irradiance is a reflection of such factors. The study conducted by Wang et al., (2012) made application of Equation (2.5) in an attempt to estimate the horizontal solar irradiance time series values. The clear shortcoming of Equation (2.5) is neglecting the account of random (stochastic) component. It also follows, from Equation (2.3), the relationship between the three solar irradiance components on a horizontal surface is given by the following equation:

$$I_g = I_b \cos\theta_z + I_d. \tag{2.6}$$

Equation (2.6) is fundamental to the calibration of solar instrumentation and implies that the vertical component of the direct beam is equal to the difference between the total and diffuse sky radiation. For tilted surfaces, Equation (2.6) can be adjusted to take the following form:

$$I_g = I_b \cos\theta_z + \; R_d D + R, \tag{2.7}$$

where θ_z is the incidence angle with respect to the normal of the tilted surface, and R_d is a conversion factor that accounts for the reduction of the sky view factor and anisotropic scattering, and R is radiation reflected from the ground that is intercepted by the tilted surface (Iqbal, 1983).

2.3 Clearness Indices: Effects of Atmosphere

The *clearness index*, denoted by K_C, generally refers to the ratio of the actual irradiance value on the ground to the extra-terrestrial beam value at the top of the atmosphere. The ratio of total irradiance on a horizontal surface, to the extra-terrestrial on a horizontal surface $I_{0,h}$ is called *clearness index for global total hemispherical*, denoted by K_G, i.e. the portion of extra-terrestrial irradiance reaching the earth's surface (see e.g. Badescu, 2008):

$$K_G = \frac{I_G}{I_{0,h}} = \frac{I_G}{I_0 \cos\theta_z}. \tag{2.8}$$

The parameter K_G is commonly used as an indicator of the relative clearness of the atmosphere and can be calculated for each daylight unit period. In general, when the atmosphere is clear, a smaller fraction of the irradiance is scattered. Basically, a low clearness index implies, for instance, a small portion of radiation reaching the surface, which reflects an overcast weather situation and hence a high diffuse fraction. On the other hand, a high clearness index indicates a clear sky weather pattern, with small diffuse radiation and hence a low diffuse fraction. The intermediate values of clearness index indicate a partly-cloudy sky conditions.

Similarly, the other two indices relating to direct beam and diffuse irradiance components (i.e. degree of cloudiness according to direct and diffuse components), are respectively given by:

$$K_b = I_b/I_0 \text{ and } K_d = I_d/I_{0,h}. \tag{2.9}$$

Moreover, at the short term, the behaviour of solar radiation is mainly ruled by the parameters such as frequency of the clouds and water vapour among others. Thus, the actual solar irradiance can be considered as the sum of two components: deterministic and stochastic. Therefore, this means that in order to isolate the stochastic component, it is necessary to normalize the irradiance value to extra-terrestrial beam value, thus accounting for the transparency of the atmosphere. That is, the ratio of the actual irradiance on the ground to that initially available at the top of the atmosphere can be calculated and presented as the degree of cloudiness indicator in the short term. This rational quantity is referred to as *instantaneous clearness index* and is required to focus on the analysis of fluctuations in solar irradiance. These indices can also be defined for the irradiation by integrating the instantaneous irradiance values over a given time interval.

2.4 Classification of weather (days)

There are two essential, generally accepted, methods (called data filters) for classifying days on the basis of the magnitude of a related parameter. These parameters are clearness index and degree of cloudiness. According to Badescu (2008), Barbaro et al. (1981), the clearness on a particular day may be judged in terms of the degree of cloudiness, both in octas and tenths as shown in Table 2.1. On the other hand, Iqbal (1983) proposed that the magnitude of the daily clearness can be measured by the so called clearness index K_T (the ratio of the solar global to the extra-terrestrial solar irradiation) to indicate the degree of cloudiness (see Table 2.2). The two methods are reported in the tables below.

Table 2.1: Classification according to cloud cover.

Day type	Octas	Tenths
Clear	0 – 2	0 - 3
Partially cloudy	3 – 5	4 - 7
Cloudy	6 – 8	8 - 10

Table 2.2: Classification according to clearness index.

Day type	Kt
Clear	$0.7 \leq Kt < 0.9$
Partially Cloudy	$0.3 \leq Kt < 0.7$
Cloudy	$0.0 \leq Kt < 0.3$

2.5 Photovoltaic (PV) System Design and Optimization

The variability of solar resource over time has a considerable impact on the solar system design. A PV array's performance is dependent on the weather, specifically on the daily levels of available solar irradiation. A series of statistical algorithms utilizing available data on solar irradiance levels at a given site are critical to the design process. Such algorithms are useful for managing the energy storage and demand by the load, which is powered by a PV array and a battery bank. The result is a statistical prediction of the PV system's performance.

The three main blocks in energy harvesting and management are the *harvesting source*, the *load* and the *harvesting system*. *Harvesting Source* refers to any available harvesting technology, such as a solar cell and a wind turbine, amongst others, which extracts energy from the environment. The *load* refers to the energy consuming activity being supported. *Harvesting system* refers to the system designed specifically to support a variable load from a variable energy-harvesting source when the there is a mismatch between the power supply levels and the consumption levels of the load. Kansal et al. (2007) presented the diagram in Figure 2.3 illustrating energy harvesting from the environment.

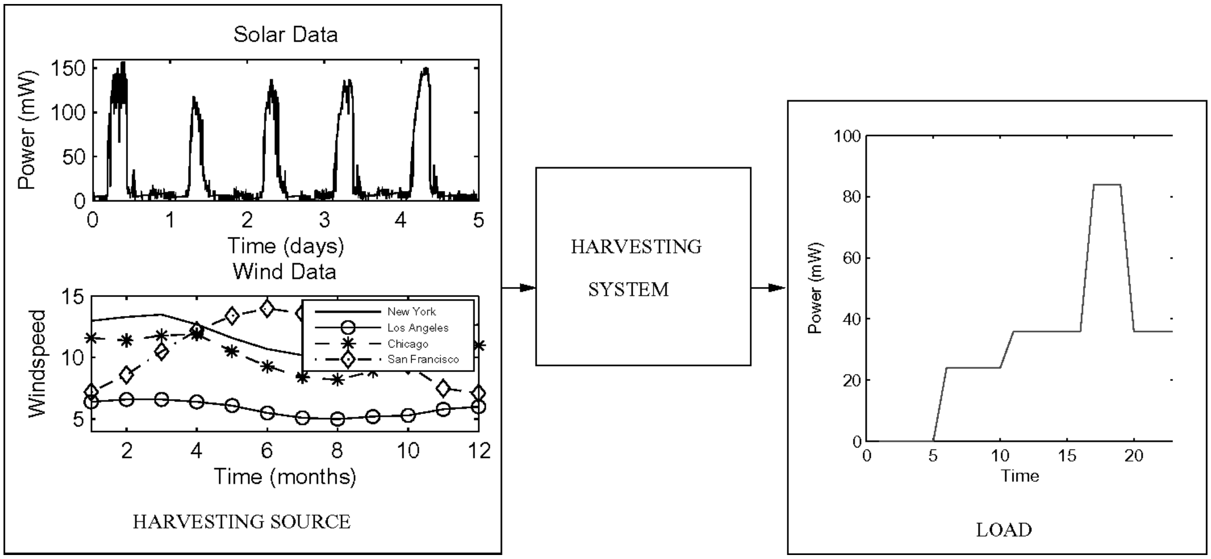

Figure 2.3: Energy harvesting from the environment with the load showing different power levels.

Energy-neutral operation and maximum performance

In energy harvesting and power-management, design considerations such as *energy-neutral operation* and *maximum performance* are critical to energy system sizing and optimization
(see e.g. Kansal et al., 2006). Such considerations depend on the system's total harvested energy. Optimal energy usage and battery sizing are also challenging issues in the process. The three main components in energy harvesting process are the harvesting source (e.g. solar cell), load and harvesting system. The whole idea is to ensure a consistent and sufficient power supply from the energy conversion system to constantly meet the energy demands of the consumption system. We elaborate on these concepts below.

Energy-neutral operation: For efficient operation the system must obviously operate such that the energy demanded by the load is continuously met or exceeded by the energy harvested. If

$P_s(t)$ is the power output from the energy source and $P_c(t)$ the consumption by the load at time instant t, then the fundamental requirement for energy-neutral operation is:

$$P_s(t) \geq P_c(t), \;\; \forall\, t. \tag{2.10}$$

The inequality in Equation (2.10) is based on the assumption of the harvesting system with no energy storage facility, i.e. the system in which the energy is directly used by the load. Therefore, this means that the excess energy is leaked or wasted. Otherwise, we have a system with ideal energy buffer**.** For such a system, there is no energy leakage, no charging inefficiency and no capacity limit. Therefore, the following inequality should be satisfied:

$$\int_0^T P_c(t)dt \leq \int_0^T P_s(t)dt + B_0, \quad \forall\, T \geq 0, \tag{2.11}$$

where B_0 is the initial energy stored in the ideal energy buffer. Again we have another case of harvesting system with non-ideal energy buffer (e.g. battery). In this system there is leakage, charging inefficiency as well as storage limits. To describe such a system, we define a rectifier function as follows:

$$[x]^+ = \begin{cases} x, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

For this particular system, a necessary and sufficient condition without the energy buffer limit is mathematically described as follows:

$$B_0 + \eta \int_0^T [P_s(t) - P_c(t)]^+ dt \leq \int_0^T [P_c(t) - P_s(t)]^+ dt + \int_0^T P_{leak}(t)dt \geq 0 \tag{2.12}$$

An additional constraint imposing a sufficient condition for the energy buffer limit constraint is:

$$B_0 + \eta \int_0^T [P_s(t) - P_c(t)]^+ dt \leq \int_0^T [P_c(t) - P_s(t)]^+ dt + \int_0^T P_{leak}(t)dt \leq B. \tag{2.13}$$

Maximum performance: The system must also be ensuring the maximum performance level that can be supported in a given harvesting environment. This may depend, for example, on the efficiency of the system hardware components, whose time to failure may be explained by an exponential random variable with mean λ.

PV-system Battery Sizing

Let us **s**uppose that a photo-voltaic system is to be installed at a particular site. To describe the operation of the system, the long-term energy balance is generally considered between the energy generated by the Photovoltaic (PV) array, the energy consumed by the load, and the energy stored in a battery. Let us consider a time interval of n_D days in which a system is required to meet the energy user demand and suppose that we experience a constant daily solar irradiation $(H_{0,D})$ in each day (i.e. there are no day-to-day variation of solar irradiation) incident on the plane of the array. If the energy demanded or consumed by the load in one day is C, then according to Arun et al. (2006), the energy required to power the load would be supplied by an array of size:

$$P_0 = \frac{C}{H_{0,D}} \tag{2.14}$$

The array size is usually expressed as a dimensionless multiple of the parameter K_A (see e.g. Egido and Lorenzo, 1992). K_A is referred to as the *solar-to-load ratio*, of the array size, given by Equation (2.14), required to consistently supply the load during the average irradiation (see e.g. Klein and Beckman, 1987):

$$P_0 = K_A \frac{C}{H_{0,D}}. \tag{2.15}$$

If we assume the situation when the daily solar irradiation is equal to H_D, below the average value of $H_{0,D}$. During this climatic cycle, the energy storage device (battery) has to cover the daily mismatch between the energy supply and demand. Therefore, to maintain a continuous electricity supply to the load, the required battery size B, in energy units, must satisfy:

$$B \geq n_D(C - P_0 H_D). \tag{2.16}$$

If the battery size B is replaced by the days of storage $K_S = B/C,$ the condition given by inequality (2.16) for continuity of supply can be written as

$$\frac{1}{n_D}K_S + \frac{C}{H_{0,D}}K_A \geq 1. \tag{2.17}$$

It is interesting to note that the inequality in Equation (2.17) is a family of straight lines with input variable K_S and output variable K_A. This represents a principal starting point for the construction of the sizing curve, based for the moment, on a single climatic cycle. The slope and the intercept of Equation (2.17) are respectively given by $m_g = -H_{0,D}K_S/n_D C$ and $1/C$.

Now, the points $(K_S, K_A) \in \Re_S$ (i.e. a shaded region) on a Cartesian plane, represent all system configurations that comply with the inequality in Equation (2.17). This method can be extended to describe real life situations with the accurate *analysis* and *simulation* of time series data.

System Sizing by Net Power Flow and Energy Balance

If P_B is the energy storage capacity of the system, P the input power from any source (e.g. photovoltaic panel), C the demand or consumption power, η_c the charging efficiency and η_d the discharging efficiency at any time point t. Then according to Arun et al. (2006), the storage rate at any time instant t is given by:

$$\frac{dP_B}{dt} = (P - C)f, \tag{2.18}$$

where

$$f = \begin{cases} \eta_c & \text{if } P \geq C \\ 1/\eta_d & \text{if } P < 0\,. \end{cases}$$

It should be noted that $P(t)$ is not the probability measure in these system sizing matters. Then the conservation energy P_B at any instant t would be given by:

$$P_B(t) = \begin{cases} P_B(t-\Delta t) + \int_{t-\Delta t}^{t} \big(P(t) - C(t)\big) f dt & \text{for continuous case} \\ P_B(t-\Delta t) + \big(P(t) - C(t)\big) f \Delta t & \text{for disctrete case} \end{cases} \tag{2.19}$$

At the instants when $P(t) \geq C(t)$, the energy surplus would be used for charging the battery. But if at any time instant t, we have $P(t) < C(t)$, then the battery makes up the energy gap. It is assumed that η_c and η_d are constant and that the variation in the battery energy with time takes place without any self-discharge losses. Given the expected load time series $C(t)$ for the site, values of η_c and η_d as well as the resource data time series $Y(t)$ in the form of global solar insolation at the specified times $t = 0, 1, \ldots, T$, it is possible to determine the minimum capacity of the power generator (P) and related battery bank (B) rating for meeting the demands of the specified load.

For obtaining the minimum generator requirement, a numerical search is performed to obtain that constant minimum value of P satisfying the following conditions:

$$\begin{gathered} P(t) \geq 0 \\ P(0) = P(T). \end{gathered} \tag{2.20}$$

The latter condition is called the repeatability condition and maintains that there is no net energy drawn from the battery for the time period considered. It is assumed that the load is recurring in the same pattern after time T. Therefore the required battery bank capacity (B) would be obtained as:

$$B = \frac{\max\{P(t)\}}{\text{DOD}}, \tag{2.21}$$

where DOD is the allowable *depth of discharge* of the battery, suitably assumed. This provides the value of the minimum possible generator capacity (P_{min}) and the corresponding sizing of the battery bank (B). It is of interest from a design perspective to identify the various feasible combinations for the generator and the storage which forms the design space for the system.

Sizing curve and design space for a cumulative energy balance

Time-series modelling includes the area of stochastic prediction and the optimal prediction of a signal sample (in a minimum mean-square sense), given a finite number of past samples. All these models are based on simplifying statistical assumptions, about the measured data.

As an example, Figure 2.4 illustrates:

(a) Daily solar radiation variation incident at a particular station during the period of 1989–1990, showing the dominant climatic cycle extending from 1st December 1989 to 7th January 1990.

(b) Cumulative energy balance (energy taken out of the battery or consumed) for a system design based on the average daily radiation in December. The average daily irradiation H_0 (shown by the dash-dot line) is the long mean value for December. Assuming the availability of a reliable model-simulated time series by which the system design may be supported sizing would then be a simpler matter. It should be noted that such sizing method considers a harvesting system with no energy consumer operating concurrently with the harvesting process. The same sizing scenario may be applied for short term battery sizing (see Markvart et al., 2006).

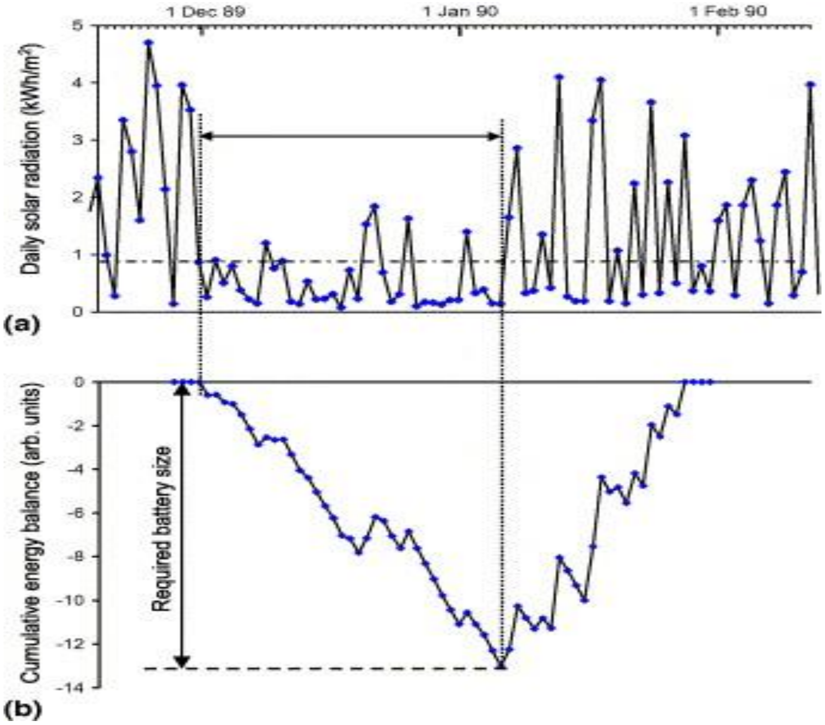

Figure 2.4: Time series curve illustrating PV-battery system sizing. *Source*: http://www.elsivier.com/locate/solener.

2.6 Statistical Models for Irradiance and Some with Physical Quantities

The amount of solar energy that reaches the earth in one hour is sufficient to supply the world's energy needs for one year and harvesting this energy efficiently is a huge challenge (Srivastava and Pandey, 2013). For such reason, it is therefore essential that some reliable mathematical models be developed to estimate the solar radiation for places where measurements are not carried out and for places where measurement records are not available.

The two common approaches that are used to study the behaviour of solar radiation on the earth's surface are Physical Modelling and Statistical Modelling. *Physical Modelling* studies the physical processes occurring in the atmosphere and influencing solar radiation. Finally, the radiation on the surface depends on the absorption and scattering processes in the atmosphere. This approach is exclusively based on physical considerations and dictates models that account for the estimated solar radiation at ground level in terms of a certain number of physical parameters such as water vapour content, dust, aerosols, clouds and cloud types, etc. The review of literature on the estimation of solar irradiance also shows that various empirical models for different geographical and meteorological conditions have been developed for estimating the monthly average daily global solar radiation on a horizontal surface (see e.g. Ulgen and Hepbasli, 2004). In their study, Ulgen and Hepbasli (2004) compared some existing models used for estimating the monthly average daily global solar radiation on a horizontal surface for some three big cities in Turkey. The outcome of this study reveal that empirical correlations are a reasonably good estimation for global radiation and through comparing the previously reported results and some two newly proposed models' results, it was found that the present models make better predictions than other previous models on the basis of various statistical measures such as MBE and RSME amongst others. These are a first order regression model and a third order polynomial model.

Statistical Solar Modelling is another important tool used to reach immediate goals in solar energy conversion. This methodology is very wide. However, the focus of this study has largely been on assessing solar irradiance time series data and the application of sophisticated time series data modelling techniques. Meteorological variables such as daylight length (sunshine duration), air temperature and relative humidity have been used as key factors in correlation models used for estimating the monthly daily global solar irradiation. A correlation making use of irradiance components and clearness index has also been established in an effort to estimate diffuse irradiance. In the next subsections, we briefly elaborate on such models of which some of them also incorporate some geographical quantities in the predictors' vector of the model. The first of these models (Angstrom equation), from which other models were derived through various modifications, is linear in nature. At the end of the section we also elaborate on various irradiance time series models that have been considered by other researchers in their attempts to model the stochastic variations of irradiance time series data.

Linear Models

Angstrom-type equation (estimation through sunshine duration)

The first ever correlation model relating solar radiation and sunshine duration was proposed by Angstrom (1924) and further modified by Prescott (1940), (see e.g. Tymvios et al., 2005). In this model, a ratio of the average day hourly global irradiation (H), to the corresponding value on a completely clear day (H_0), and the ratio of the average daily sunshine duration (S) to the maximum possible sunshine duration, S_0, are related through the linear Equation (2.22), (see e.g. Almorox et al., 2004; Srivastava and Pandey, 2013).

$$\frac{H}{H_0} = a + b\left(\frac{S}{S_0}\right), \tag{2.22}$$

The constants a and b are determined regression parameters that can be estimated for different locations using simple linear regression. This linear relationship is also known as the Angstrom–Prescott Equation, named after the proposal by Prescott (1940) that the average global irradiation on a clear day should be replaced with the extra-terrestrial intensity values to put the equation in a more convenient for the clear sky global irradiance might not be determined exactly. From Equation (2.22), a unique model for each month is then estimated from the measurements obtained for that particular month (see e.g. Ulgen and Hepbasli, 2004).

Estimation through air temperature and relative humidity

In this model, the regressor comprises the ratio of the measured day temperature (T) to the maximum possible temperature (T_0), i.e. the hottest air temperature reported on earth, to predict the ratio of average day hourly global solar radiation to its corresponding value on a completely clear day.

$$\frac{H}{H_0} = a + b\left(\frac{T}{T_0}\right). \tag{2.23}$$

Similarly, a correlation model comprising the ratio of the measured relative humidity (M) to the maximum possible relative humidity (M_0) is given by:

$$\frac{H}{H_0} = a + b\left(\frac{M}{M_0}\right). \tag{2.24}$$

Estimation of diffuse fraction through the clearness index

For the stations where only measurements of the global irradiance (I_g) may be available, a correlation model for estimating the diffuse fraction (I_d) when it is not known has been suggested. This model correlates the diffuse fraction with the clearness index and is developed from the measured values of both total and diffuse irradiance on a horizontal surface over a certain period of time. The ratios $K_g = I_g/I_{0,h}$ (ratio of global irradiance to extra-terrestrial horizontal irradiance) and $f_d = I_d/I_g$ (ratio of diffuse irradiance to global irradiance) obtained for each daylight unit period are related through the following equation,

$$K_g = \alpha_0 + \alpha_1 f_d. \tag{2.25}$$

For easy modelling purposes, Equation (2.25) may be justified for a binary random variable defined on $[0, 1]$, to take the following form,

$$f_d = \frac{1}{1+e^{\alpha_0+\alpha_1 K_g}}. \tag{2.26}$$

This is called a logistic function, used for estimating proportions. There are various methods for performing the fit. One common method is to transform Equation (2.26) into a linear equation in α_0 and α_1, as follows:

$$\ln\left(\frac{1-f_{d,i}}{f_{d,i}}\right) = \alpha_0 + \alpha_1 K_{g,i} \ . \tag{2.27}$$

The model parameters α_0 and α_1 are then estimated by using an iterative procedure such as the Newton Raphson algorithm. Many linear relationships exist indeed between solar variables and meteorological factors and also among irradiance components themselves, e.g. global solar irradiance and diffuse fraction. The strength of correlation between these variables will of course depend on the sky conditions, e.g. on overcast days, I_g and I_d are almost equal.

Polynomial Models

According to Ulgen and Hepbasli (2004), Angstrom–type equation has been further revised and modified by Samuel (1991) and Zabara (1986) to higher degree polynomial functions, e.g. quadratic and third degree functions. The proposed polynomial regression models are given by

$$\frac{H}{H_0} = a + b\left(\frac{S}{S_0}\right) + c\left(\frac{S}{S_0}\right)^2 \ \text{ and } \frac{H}{H_0} = a + b\left(\frac{S}{S_0}\right) + c\left(\frac{S}{S_0}\right)^2 + d\left(\frac{S}{S_0}\right)^3 . \tag{2.28}$$

According to Zabara (1986), the parameters of the modified Angstrom model, a and b, can be correlated with the maximum possible sunshine duration (S_0) and daylight length (S) as a third order function as follows,

$$\begin{aligned} &\frac{H}{H_0} = 0.14 + 2.52\left(\frac{S}{S_0}\right) + 3.71\left(\frac{S}{S_0}\right)^2 + 2.24\left(\frac{S}{S_0}\right)^3 \\ &a = 0.395 - 1.247\left(\frac{S}{S_0}\right) + 2.68\left(\frac{S}{S_0}\right)^2 - 1.674\left(\frac{S}{S_0}\right)^3 \\ &b = 0.395 + 1.384\left(\frac{S}{S_0}\right) - 3.249\left(\frac{S}{S_0}\right)^2 + 2.055\left(\frac{S}{S_0}\right)^3 . \end{aligned} \tag{2.29}$$

Angular models and other models

Also, from the Angstrom-type equation, another class of models called angular models has been developed. One such model, proposed by Gopinathan (1988), makes use of the cosine of the latitude (ϕ), elevation (h) and percentage of possible sunshine for any location around the world to estimate the parameters a and b as follows,

$$a = -0.309 + 0.539\cos\phi - 0.0693h + 0.29\left(\frac{S}{S_0}\right)$$
$$b = 1.527 - 1.027\cos\phi + 0.0926h - 0.359\left(\frac{S}{S_0}\right). \qquad (2.30)$$

According to Glover and McCulloch (1958), a good estimation may be achieved through Equation (2.31) at latitudes of $\phi < 60^0$.

$$\frac{H}{H_0} = 0.29\cos\phi + 0.52\left(\frac{S}{S_0}\right),\ \phi < 60^0. \qquad (2.31)$$

A similar model incorporating latitude (ϕ) was formulated by Raja and Twidell (1990) as follows:

$$\frac{H}{H_0} = 0.388\cos\phi + 0.367\left(\frac{S}{S_0}\right). \qquad (2.32)$$

Among other models we have one incorporating the logarithmic term, proposed by Newland (1988) and given by:

$$\frac{H}{H_0} = 0.34 + 0.4\left(\frac{S}{S_0}\right) + 0.17\log\left(\frac{S}{S_0}\right). \qquad (2.33)$$

In addition to (S/S_0), only the altitude of the site (h) was taken into account with the values of a and b adjusting, (Gopinathan, 1988) to:

$$a = 0.265 + 0.07h - 0.0135\left(\frac{S}{S_0}\right) \text{ and } b = 0.401 - 0.108h - 0.325\left(\frac{S}{S_0}\right). \qquad (2.34)$$

According to Dogniaux and Lemoine (1983), the regression coefficients a and b can be given as linear functions of the latitude (ϕ) in average and on the monthly basis, as follows:

$$a = 0.37022 - 0.00313\phi \text{ and } b = 0.32029 - 0.00506\phi. \tag{2.35}$$

2.7 Forecasting Models for Irradiance on Various Time Scales

Literature reveals that model (2.2) has been used by some researchers in the attempt to model and forecast the irradiance time series values (see e.g. Wang et al., 2012). In this study, the surface irradiance measurements for four consecutive days on hourly scales (from 9th to 12th of March 2010) were examined to assess the difference between the surface measured values (I_0) and their extra-terrestrial counterparts (I_g), i.e. $D_I = I_0 - I_g$, referred to as the solar irradiance difference. Apparently, as revealed by figures, a major drawback of such a model is that it only takes into consideration geographic quantities, ignoring the account of the influences by the random and nonlinear characteristics of solar irradiance under changeable weather conditions. As a consequence, the solar irradiance series values have been overestimated and/or underestimated (see e.g. Wang et al., 2012, Figure 1).

The results showed that the solar irradiance difference is less variable on clear sky (sunny) days (e.g. 11th and 12th of March 2010), larger and has more inflections on cloudy or overcast days (e.g. 9th and 10th of March 2010). This variation is apparently related to the weather conditions unfavourable for the maximum amount of energy to be received on the surface. The analysis of the variation related to different weather conditions can be useful for extracting more information from the measured values of surface solar irradiance and their extra-terrestrial counterparts, by finding and selecting suitable climatic parameters such as ambient temperature and relative humidity amongst others. Such parameters can be reflective of these correlative variation characteristics, i.e. they can reflect the changes of irradiance and be considered as the input of other time series forecasting models such as an Artificial Neural Network (ANN) model.

Artificial neural networks (ANNs) methodology

Artificial neural networks are a class of distinct mathematical models originally motivated by the information processing in biological neural networks, and have found applications in forecasting tasks and modelling nonlinear functions, e.g. solar radiation forecasting (see e.g. Khatib et al., 2012; Wang et al., 2012 and Paoli et al., 2009). ANN's learn from sample data by constructing an input-output map without explicit analytical expression of the model equation, thus modelling complex relationships between inputs and outputs. They can be used to model any actual system by changing its connection weights based on external or internal information that flows through the network during its learning from existing sample data.

In another study by Martin et al. (2010), Neural Networks (NN) method has been used in an attempt to forecast half daily values of solar irradiance for the next three days at different solar power stations. Two further methods namely, Autoregressive (AR) model and Adaptative-network-based fuzzy inference system (ANFIS) models were used in comparison with NN method. According to these authors, ANFIS models are a class of neural networks which are functionally equivalent to fuzzy logic inference systems. Due to non-stationary behaviour of half daily global solar irradiance time series, it was necessary to transform data to two new variables namely, clearness index (the ratio between ground measured global solar irradiance and extraterrestrial solar irradiance) and lost component (the difference between extraterrestrial solar irradiance and ground measured global solar irradiance).The accuracy of the three models to forecast half daily values of solar irradiance was measured on the basis of root mean square error (RMSE). Neural network and ANFIS models with lost component as input were found to be the better approaches except at one station where clearness index time series is easier to simulate by models. The results also showed that the clearness index time series obtains better results in models of lower order compared to lost component. AR models from time series shows higher uncertainty than nonlinear models. The clear disadvantage of AR models given by these researchers is the common big size of the input vectors of parameters, deteriorating parsimony, e.g. AR(p) models of order up to $p = 10$.

Forecasting factors of the ANN model are selected from the two categories of historical data: *solar irradiance* itself and the *meteorological parameters* related to solar irradiance. According to Wang et al. (2012), the model input vector of historical irradiance data, Y_{t-1} $(i = 1, \dots, k)$, can be shown as follows,

$$V = [Y_{t-1}, Y_{t-2}, \dots, Y_{t-k}]. \tag{2.36}$$

This study revealed that the information about meteorological factors reflected by the difference between the surface irradiance and extra-terrestrial irradiance is useful for the development of the ANN model. Such factors can be used as predictors in the model and if incorporated in appropriate forms, they can make forecasting even more precise. Some research studies revealed that the derivative index is useful for describing the variation tendency of the difference D_I. Because D_I is closely related to weather variations, so are the derivatives of D_I. Further studies showed that the three derivatives (dD_I/dt, d^2D_I/dt^2, d^3D_I/dt^3) are all positively correlated with the intensity variations of surface irradiance (I_g). The third order derivative (d^3D_I/dt^3), being greater than the 1st and 2nd order is appropriate for describing rapid and violent fluctuations of the weather. In order to get a more significant, clear and simple description for different weather conditions of one day, the maximum value of d^3D_I/dt^3, denoted by TOD_{max}, is recommended as an appropriate index for changeable weather characteristics.

Another key factor is the shape difference between I_g and I_0, which is also closely related to changeable weather conditions. Hence, clear comparisons of the shape difference between I_g and I_0, can be made using the normalized irradiance values, $I_{gN,i}$ and $I_{0N,i} \in [0,1]$, in Equation (2.37), i.e. the values of each series (of length k) are divided by their maximum to eliminate the impact of different amplitudes of the actual irradiance values.

$$I_{gN,i} = \frac{I_{g,i}}{\max_{i=1,2,\dots,k}\{I_{g,i}\}} \text{ and } I_{0N,i} = \frac{I_{0,i}}{\max_{i=1,2,\dots,k}\{I_{0,i}\}} \ . \tag{2.37}$$

The normalized discrete difference (NDD) of solar irradiance is defined in Equation (2.38) as a specific index for describing the difference between I_g and I_0. This method gave better results since the differences were significantly reduced (see Wang et al., 2012, Figure 3). The NDD values computed are referred to as the NDD index. The NDD index can be useful for measuring different weather conditions.

$$\mathrm{NDD} = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(I_{0N,i} - I_{gN,i})^2}\ . \tag{2.38}$$

In addition to previously discussed predictors, other variables such as day number of the year $N \in [1,365]$, average surface irradiance $\bar{I}_g$ and the average day ambient temperature $\bar{T}$ may also be included in the input vector. Therefore, the final input vector of the new ANN forecasting model may be composed of five components as shown below:

$$I_{ANN} = \left[\bar{I}_g, TOD_{max}, NDD, \bar{T}, N\right]. \tag{2.39}$$

The ANN model consists of an interconnected group of neurons, referred to as the endogenous entries and output variables in the processing stages of computation. The three processing stages called layers are input layer, hidden layer and output layer. Usually there are more than one hidden layer each with a certain number of neurons, e.g. there are two hidden layers in the ANN model of the above mentioned source with p and q number of neurons in hidden layer1 and hidden layer 2 respectively. The input layer of ANN consists of the input vector I_{ANN}, given by Equation (2.39). The output vector consists of the forecasted values for the next couple of periods. The output of the network z_i can be represented by

$$z_i = \sum_{j=1}^{n} w_{ij}x_{ij} + \theta_i, \tag{2.40}$$

where x_{ij} is the incoming signal from the jth neuron (at the input layer), w_{ij} is the weight on the connection directed from neuron j to neuron i (at the hidden layer) and θ_i is the bias of neuron i. After each z_i is calculated, an activation function is applied to modify it. The activation function is typically a bounded monotonic function such as the standard logistic sigmoidal function defined by $f(z_i) = 1/(1 + e^{-z_i})$. According to Jiang (2008), the number of neurons of the hidden layer can be expressed as $m = (p + q)^{0.5} + a$, where m is the number of neurons of the hidden layer, p is the number of neurons of the input layer, q is the number of neurons of the output layer and a is a constant from 1 to 10 (Jiao, 1990).

The output vector consisted of 24 hours ahead forecasts which represent the surface irradiance of the 24 hours of the next day (season). Two different series for cloudy days and sunny days separately were modelled. For comparison purposes, the conventional model called ANN-Hybrid Discrete Continuum (ANN-HDS) was developed on the same dimensions and time horizons. The error statistical indicators such as MAPE, RMSE and MABE were used in measuring the forecast accuracy. The irradiance was forecast on a time scale of 24–72 hours ahead, which is considered short term forecasting. The results show the ANN models give reasonably good forecasts with the suggested input vector of statistical parameters.

ANN with time series pre-processing on a daily time scale

Nonlinear variability in the data is usually dealt with by the application of neural networks. Another application of neural networks has been made by Paoli et al. (2009) in the prediction of daily global solar radiation on a horizontal surface. In this study, a methodology making use of ad-hoc time series pre-processing and a Multi-Layer Perceptron has been developed for predicting daily global solar radiation on a daily horizon. The global solar radiation time series data collected on a daily basis from the solar meteorological station of Ajaccio France, located at 41°55'N, 8°44'E, from Jan 1971 to Dec 1989 was examined in this study. To quantify the annual periodicity, the original series (H) was divided by daily extra-terrestrial radiation (H_0) to form a new series X_t (called *index clarity*) defined as follows:

$$X_t = \frac{H}{H_0}. \tag{2.41}$$

Despite this pre-treatment, Fisher’s test indicated that the seasonality was not optimal. In their case, they found that it led to a new seasonality which is difficult to model. The optimal seasonality was then achieved by using a ratio to moving average after a ratio to trend method (using H_0) to correct rigid seasonalities. The latter can be applied when there is no analytical expression of the trend. In the case of flexible seasonality, i.e. random in amplitude or period, the filtering techniques by successive moving averages are recommended, i.e.:

$$Y_t = \frac{X_t}{\frac{1}{2m+1}\sum_{i=-m}^{m} X_{t+i}} \,. \tag{2.42}$$

In this case, $2m + 1 = 365$ days suggests that $m = 182$. Therefore, to complete the process, 365 seasonal factors (S_t) were used. These are indeed the coefficients that get rid of the rigid seasonality by moving average ratio given by Equation (2.42). The transition coefficients ($N = 18$, number of years of history) and the average coefficients of the regular 365 days are given as $S_t = (1/N)\sum_{j=1}^{N} Y_{t,j}$ and $\bar{S}_t = (1/365)\sum_{j=1}^{365} S_t$ respectively. Then the final seasonal factors are given by

$$S_t^* = \frac{S_t}{\bar{S}_t} \,. \tag{2.43}$$

Hence, it follows a new series, seasonally adjusted, that represents only the stochastic component of global radiation:

$$S_t^{corr} = \frac{X_t}{S_t^*} \,. \tag{2.44}$$

This particular method provided better results than the other methods, including an ARIMA model, on the basis of RMSE measure.

Lucheroni model

According to Huang et al. (2011), this is one of the key approaches to modelling global solar radiation on short time basis (e.g. hourly), with its origin in biophysics (Lucheroni, 2007). This model is given in the discretized version of the model for the deseasonalized solar radiation time series R_t, as follows:

$$\begin{gathered} R_{t+1} = R_t + \upsilon_t \Delta t + o_t \\ \upsilon_{t+1} = \upsilon_t + [\kappa(\upsilon_t + R_t) - \lambda(3R_t^2\upsilon_t + R_t^3) - \epsilon\upsilon_t - \gamma R_t - b].\frac{\Delta t}{\epsilon} + a_t, \end{gathered} \tag{2.45}$$

where υ_t is the derivative of R_t, o_t and a_t are noise terms at time t and Δt is the time step. The aim of the model (2.45) is to exploit the fact that the current value of υ_t is useful to predict the future value, R_{t+1}. The parameters κ, λ, ϵ, γ and b can be estimated using the method of ordinary least squares. This model was used to fit the same data from Mildura, together with an AR(2) model. The results indicated that this model actually described effectively the pattern of the deseasonalized data, with its ability to capture the magnitude of peaks and troughs almost perfectly. However, a disadvantage of this model is its inability to perform well when residuals are decreasing.

Coupled Autoregressive and Dynamical System (CARDS) model

From a similar study conducted by Huang et al. (2011), it was concluded that the Lucheroni model performed poorly in comparison with the AR(2) model, with a high margin of forecast error. To improve on this, the Combined Autoregressive Dynamical System (CARDS) forecasting method on a short time scale was then introduced. This model is composed of a mixture of both the AR(2) and the Lucheroni model, combined to develop a method with a better forecasting profile. This new combination model is defined by:

$$f_t = M_t + \nabla_t^1, \tag{2.46}$$

where M_t is the prediction obtained from the model at time t, $\nabla_t^1 = R_t - R_{t-1}$ and f_t, a notion of a "fixed component", intended to replace M_t. However, not all predictions from the combination model are replaced by the fixed component values but only under certain conditions that some predictions may be replaced. The applicability of the model given by Equation (2.46) has been demonstrated for one-step-ahead forecasting on hourly and sub-hourly time scales. The results showed that the CARDS model follows the variation in the observed data series better and hence improves forecasting.

Forecasting solar irradiance with ARIMA models

An Autoregressive Integrated Moving Average (ARIMA) model, discussed in the following chapters, has also been used by to model and forecast solar irradiance on an hourly scale (see e.g. Dazhi et al., 2012). In this study, three forecasting methods taking into account the effect of cloud cover were proposed using three types of solar radiation data as input parameters, namely, global horizontal irradiance (GHI), diffuse horizontal irradiance (DHI), direct normal irradiance (DNI). The first method directly uses GHI to forecast next hour GHI through additive seasonal decomposition followed by an Auto-Regressive Integrated Moving Average (ARIMA) model. The second method forecasts DHI and DNI separately using an additive seasonal decomposition followed by an ARIMA model. The two forecasts are then combined to predict GHI using an atmospheric model. The third method considers cloud cover effects. An ARIMA model was used to predict cloud transients. GHI at different zenith angles and under different cloud cover conditions was constructed using nonlinear regression.

The three methods were tested using data from two different weather stations and it was found that the forecasts using cloud cover information can improve the forecast accuracy. However, it is believed that cloud cover can increase the forecast accuracy only if the data set is sufficiently accurate to represent the actual situation, i.e. if the hourly solar irradiance values do not deviate significantly from actual values for partly cloudy skies conditions. Under partly cloudy skies conditions, it is recommended that the sampling frequency of cloud cover is increased for better reflection of the true values for partly cloudy skies conditions. Furthermore, ARIMA modelling was used as it is believed that this approach can deal with both stationary and non-stationary time series, and can also be used with integrated and moving average process orders. In our study we also use ARIMA as one of the approaches, however taking into account the seasonality inherent in the data as the author also acknowledges the presence of seasonality by the plots. That is, in our study we make use of SARIMA accounting for the seasonal behaviour. As it is also the case with our study, these researchers made use of the Akaike information criterion (AIC) search algorithm to search for the optimal model which fits the specific time series. The criteria used to evaluate the forecast accuracy are the mean bias error (MBE) and root mean squared error (RMSE), defined in one of the succeeding chapters.

Fourier series model for an hourly global solar irradiance

The same study by Huang et al. (2011) reveals that a time series with seasonality S_t can be described by a Fourier series model in the context of spectral (harmonic) analysis. This method has been used for modelling hourly values of global irradiance for three consecutive days in Mildura. A descriptive model of this kind for the series was given by:

$$
\begin{aligned}
S_t = \mu + \alpha_1 \cos\frac{2\pi t}{8760} \\
&+ \beta_1 \sin\frac{2\pi t}{8760} \\
&+ \alpha_2 \cos\frac{4\pi t}{8760} + \beta_2 \sin\frac{4\pi t}{8760} \\
&+ \sum_{n=1}^{3}\sum_{m=-1}^{1}\left(\alpha_{nm}.\cos\frac{2\pi(365n+m)t}{8760} + \beta_{nm}.\sin\frac{2\pi(365n+m)t}{8760}\right)
\end{aligned}
\tag{2.47}
$$

where t is the time in hours, μ is the mean of the data, α_1 and β_1 are coefficients of the yearly cycle, α_2 and β_2 of twice yearly and α_i and β_i coefficients of daily cycle and its harmonics (n = 2, 3 and n = 1) and associated beat frequencies ($m = 1$). The latter modulate the amplitude to fit the time of year (i.e. the beating of the yearly and daily cycles).

In this study, the frequencies of yearly, twice yearly, daily and twice daily cycles, were determined and coefficients of determination for each Fourier series components were computed. The yearly cycle explained a small percentage of the variance of the series, while the daily and twice daily cycles explained over 70%. This is an indication of a very strong daily cycle and a less prominent yearly cycle. From the outcome of this study, it has been found that the Fourier series alone is not enough to model global solar radiation, due to the underestimation of irradiance for some days and overestimation for others. Therefore, as part of this study, we undertake to show that the application of HCSARIMA models indeed yields better results with the irradiance data for UKZN HC Solar Station.

Chapter Summary

In chapter 2, an overview of some previous relevant research studies in the field of solar energy has been given, i.e. this chapter highlights a picture of the existing knowledge and previously attempted models for solar radiation. From this chapter, we have gained insight into how the incoming energy from the sun is influenced by meteorological quantities, e.g. clouds, and geographical quantities, e.g. surface zenith angle, as it traverses to a collection system on the ground. The solar energy from the sun is mostly influenced by air particles as it passes through the atmospheric layer to the surface. Just before reaching the surface, it also interacts with the cloud cover. These phenomena are causing reduction of the solar energy falling on the ground. Consequently, the energy incident on a solar measurement system typically comes in three forms viz., direct irradiance, diffuse irradiance and global irradiance. The original (unaffected) irradiance coming directly from the sun is known as extra-terrestrial solar irradiance, this before it passes through the atmospheric layer and undergoes reduction. The physical models have been developed in an effort to estimate solar radiation received on the ground. One such model is given by Equation (2.2). This model was built from geographical quantities relating to surface orientation where the solar collector is located. Further, we have gained an idea of measuring the sky conditions, using the clearness index, a quantity given as a ratio of the amount of energy received on the ground to the actual (original) energy amount from the sun. Such models could be useful for monitoring solar energy conversion systems. Also given in Chapter 2 are some scientific time series models that can be used in addressing the challenges arising in design and sizing of solar power systems as well as power management of such systems. The rest of the chapter gives us an overview of some attempts that have been made to estimate the solar flux on the ground. The first attempts by Liu and Jordan (1960), gave a statistical linear model correlating diffuse fraction with the clearness index. Some other related models have been reviewed. The attempts to forecast irradiance time series data have also been discussed. Examples are the application of the Artificial Neural Networks (ANN) model, the Lucheroni model, the Coupled Autoregressive and Dynamical System (CARDS) and the Fourier series model amongst others. The literature also reveals forecasting solar irradiance with meteorological parameters such as cloud cover index. However, some of these models as discussed did not perform well enough with the experimental data and therefore there is a room for further research studies.

Chapter 3
Methods of Time Series Analysis

The methods to be presented in this chapter are designed for the purpose of analysing time series observations taken at regular intervals in time. The time domain and frequency domain methods in time series analysis, will be introduced. These methods have a wide range of applications and we can mention astronomy and signal processing (see e.g. Pollock et al., 1999) amongst many others. Both methods apply to what are described as stationary or non-evolutionary time series. Such series manifest statistical properties which are invariant throughout time, so that the behaviour through one epoch is the same as it would be during any other.

3.1 Time domain analysis

In time domain analysis, we define a univariate time series as a set of random variables indexed by time, denoted by $\{Y_t\}_{t=-\infty}^{\infty}$. An observed time series $\{y_t\}_{t=1}^{n}$can be regarded as a partial realization (of sample size n) of a set of random variables $\{Y_t\}_{t=-\infty}^{\infty}$. Such a set of random variables is also called a stochastic process denoted by $\{Y_t\}_{t=1}^{n}$. Unless otherwise stated the process is often assumed to be real valued, with the values evolving in time according to some probabilistic laws.

3.1.1 Stationarity

A time series process is said to be weakly stationary if it has time invariant first and second moments, i.e. if the mean and the variance are constant and finite, whereas for a non-stationary process the mean and variance are time variant. A definition of strict stationary is given as follows,

A process Y_t is said to be strictly stationary of order n if for any n-tuple $(t_1, t_2, \dots, t_n)$, where $k \in \mathbb{Z}$, the following holds,

$$F_{Y_{t_1},\dots,\ Y_{t_n}}\left(y_{t_1},\dots,y_{t_n}\right) = F_{Y_{t_{1+k}},\dots,Y_{t_{n+k}}}\left(y_{t_{1+k}},\dots,y_{t_{n+k}}\right), \tag{3.1}$$

i.e. if the joint distribution functions of $\{Y_{t_1}, Y_{t_2}, \ldots, Y_{t_n}\}$ and $\{Y_{t_{1+k}}, Y_{t_{2+k}}, \ldots, Y_{t_{n+k}}\}$ are the same.

For a real-valued process the mean function is defined as $\mu_t = E(Y_t)$ and the variance function $\sigma_t^2 = E(Y_t - \mu_t)^2$.

A natural estimator of the process mean is the sample mean obtained from a single realization of the process, $\{Y_t\}_{t=1}^n$, and given by the following formula:

$$\bar{Y} = \frac{1}{n}\sum_{t=1}^{n} Y_t \, . \tag{3.2}$$

$E(\bar{Y}) = \frac{1}{n}\sum_{t=1}^{n} E(Y_t) = \mu$ implies that $\bar{Y}$ is an unbiased estimator of the mean.

The variance of $\bar{Y}$ is defined as follows,

$$\text{Var}(\bar{Y}) = \frac{1}{n^2}\sum_{t=1}^{n}\sum_{s=1}^{n} \text{Cov}(Y_s, Y_t) = \frac{\gamma_0}{n^2}\sum_{t=1}^{n}\sum_{s=1}^{n} \rho_{t-s} = \frac{\gamma_0}{n}\sum_{k=-(n-1)}^{k-1} \left(1 - \frac{|k|}{n}\right)\rho_k. \tag{3.3}$$

If $\lim_{n\to\infty} \sum_{k=-(n-1)}^{n-1} \left(1 - \frac{|k|}{n}\right)\rho_k \leq \infty$, then $\lim_{n\to\infty} \text{var}(\bar{Y}) \to 0$ and $\bar{Y}$ is a consistent estimator of the mean μ, i.e. $\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} Y_t = \mu$, in mean square. For this to hold, $\lim_{k\to\infty} \rho_k = 0$ is a sufficient condition.

Covariance stationary time series

Let $\{Y_t\} = \{\ldots Y_{t-1}, Y_t, Y_{t+1}, \ldots\}$ denote a sequence of random variables indexed by time t, i.e. $\{Y_t\}$ is a time series process. Then $\{Y_t\}$ is said to be covariance stationary if

$$E(Y_t) = \mu\,,\ \forall t \ \text{ and}$$

$$\text{Cov}(Y_t, Y_{t+k}) = E[(Y_t - \mu)(Y_{t+k} - \mu)] = \gamma_k, \forall t \text{ and any } k \tag{3.4}$$

For such process, $\gamma_k = \mathrm{Cov}(Y_t, Y_{t+k})$ and $\rho_k = \mathrm{Corr}(Y_t, Y_{t+k})$ are referred to as the autocovariance and autocorrelation functions, respectively.

For brevity, a covariance stationary time series can be simply called a stationary time series. The parameter γ_k is called the kth order or lag k autocovariance of $\{Y_t\}$ and a plot of γ_k against k is called the autocovariance function.

3.1.2 Autocovariance and Autocorrelation Functions

The covariance function between $Y_{t_1}$and Y_{t_2}(called autocovariance) is defined by:

$$\gamma(Y_{t_1}, Y_{t_2}) = E\big(Y_{t_1} - \mu_{t_1}\big)\big(Y_{t_2} - \mu_{t_2}\big), \tag{3.5}$$

and the correlation between $Y_{t_1}$and Y_{t_2} (called autocorrelation) as

$$\rho(Y_{t_1}, Y_{t_2}) = \frac{\gamma(Y_{t_1}, Y_{t_2})}{\sigma_{t_1}\sigma_{t_2}}. \tag{3.6}$$

Letting $t_1 = t$ and $t_2 = t + k$, the covariance between the series values Y_t and Y_{t+k} that are k time periods apart can also be expressed as:

$$\gamma_k = \gamma(Y_t, Y_{t+k}) = f(k) \,,$$

and correlation as $\rho_k = \gamma_k/\gamma_0$.

The sample autocovariance at lag k is given by:

$$\hat{\gamma}_k = \frac{1}{n}\sum_{t=1}^{n-k}(Y_t - \bar{Y})(Y_{t+k} - \bar{Y}) \,, \tag{3.7}$$

and the sample autocorrelation function (SACF) by:

$$\hat{\rho}_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0} = \frac{\sum_{t=1}^{n-k}(Y_t-\bar{Y})(Y_{t+k}-\bar{Y})}{\sum_{t=1}^{n}(Y_t-\bar{Y})^2}. \tag{3.8}$$

The standard error of the autocorrelation at lag k is based on the squared autocorrelations from all previous lags, defined as follows:

$$SE(\hat{\gamma}_k) = \sqrt{\frac{1+2\sum_{j=0}^{k-1}\hat{\gamma}_j^2}{N}}, \tag{3.9}$$

where N is the length of a series as a whole The quantity $\hat{\gamma}_0^2$ is set to 0 at lag $k = 1$, as there are no previous correlations. The standard error for a partial autocorrelation is the same at all lags and is simply given by:

$$SE(\hat{\rho}_k) = \frac{1}{\sqrt{N}}. \tag{3.10}$$

The two functions are measures of the strength of association between the current and past series values. A plot of ρ_k against k is called autocorrelation function (ACF) and gives correlation between the series values at different values of k. The hypothesis testing

$$\mathrm{H}_0: \rho_k = 0 \text{ (or } \rho_{kk} = 0)$$
$$\mathrm{H}_1: \rho_k \neq 0 \text{ (or } \rho_{kk} \neq 0)$$

is used to test for the significance of the lag k autocorrelation. By inspection of relevant plots, if $\hat{\rho}_k$(or $\hat{\rho}_{kk}$) value (represented by spike) is outside the ± 2 Standard Error lines, the null hypothesis (H_0) statement is rejected in favour of H_1.

Properties of autocovariance and autocorrelation functions:

i. $\rho_0 = 1$

ii. $|\rho_k| \leq 1$ and $|\gamma_k| \leq \gamma_0$

iii. $\gamma_k = \gamma_{-k}$ $(\rho_k = \rho_{-k})$ (3.11)

Ergodic time series: A stationary time series $\{Y_t\}$ is said to be *ergodic* if the sample moments (i.e. sample mean and sample variance) converge in probability to the population moments, i.e. if

$$\bar{Y} \overset{p}{\rightarrow} \mu\ ,\ \hat{\gamma}_k \overset{p}{\rightarrow} \gamma_k \text{ and } \hat{\rho}_k \overset{p}{\rightarrow} \rho_k \tag{3.12}$$

Partial autocorrelation function: Partial autocorrelation function (PACF) is a complementary tool which describes the partial correlation between Y_t and Y_{t+k} after adjusting for $Y_{t+1}, \dots, Y_{t+k+1}$, i.e. PACF at lag k, just as ACF, gives correlation between the series values that are k intervals apart, but accounting for the values in between. Therefore, PACF indicates which past series values are most useful in predicting future values. It is a useful tool to help identify AR(p) models and is based on estimating the sequence of AR.

The partial autocorrelation of k^{th} order is defined as follows,

$$\rho_{kk} = \text{Corr}[Y_t - \mathcal{P}(Y_t|Y_{t+1}, \dots, Y_{t+k-1}), Y_{t+k} - \mathcal{P}(Y_{t+k}|Y_{t+1}, \dots, Y_{t+k-1})] \tag{3.13}$$

where $\mathcal{P}(W|Z)$ is the best linear projection of Won Z, i.e. $\mathcal{P}(W|Z) = \sum_{WZ}\sum_{ZZ}^{-1}Z$ with $\sum_{ZZ} = \text{Var}(Z)$ the covariance matrix of regressors and $\sum_{WZ} = \text{Cov}(W, Z)$ is the matrix of covariances between W and Z.

An equivalent definition of the above is the solution to $\boldsymbol{\rho_{kk}}$ of the following system of equations:

$$P_k\boldsymbol{\rho_{kk}} = \boldsymbol{\rho_k} \tag{3.14}$$

where $P_k = \begin{pmatrix} 1 & \rho_1 & \cdots & \rho_{k-1} \\ \rho_1 & 1 & & \rho_{k-2} \\ \vdots & & \ddots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \cdots & 1 \end{pmatrix}$

$\boldsymbol{\rho_{kk}} = (\rho_{k1}, \dots, \rho_{kk})^T$ and $\boldsymbol{\rho_k} = (\rho_1, \dots, \rho_k)^T$.

These are called the Yule-Walker equations. The last coefficient, ρ_{kk}, is the partial autocorrelation of order k.

Defining $P_k^* = \begin{pmatrix} 1 & \rho_1 & \cdots & \rho_1 \\ \rho_1 & 1 & & \rho_2 \\ \vdots & & \ddots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \cdots & \rho_k \end{pmatrix}$ and using the Cramer-Rule, a general solution is given by:

$$\boldsymbol{\rho_{kk}} = \frac{|P_k^*|}{|\ P_k|}. \tag{3.15}$$

From the definition of PACF, it immediately follows that there is no difference between PACF and ACF of order one, i.e. $\rho_{11} = \rho_1$.

3.1.3 Data Transformations

Trend and seasonality

A transformation is applied to time series data either to remove trend and cycles (seasonality) or to stabilize the variance. The presence of trend in the time series leads to non-stationarity. Therefore, before attempting to use the Box-Jenkins ARIMA models, it is often worth transforming the data. In removing trend from the series Y_t the dth differencing operator is often applied to create a new stationary series Z_t, with time invariant first and second moments. For example, the first differencing applied to a series with a linear trend eliminates the trend yielding the transformed series

$$Z_t = (1 - L)^d Y_t = \nabla^d Y_t\ , \tag{3.16}$$

where d is the regular differencing operator and L is the backward shift operator, i.e. $LY_t = Y_{t-1}$.

If the seasonality of length S exists in a series, a Dth differencing will remove it, to result in the following

$$Z_t = (1 - L)_S^D Y_t = \nabla_S^D Y_t, \tag{3.17}$$

where D is the seasonal differencing operator.

To extract trend and /or cycles in a time series, symmetric moving average (MA) smoothing is generally employed. It makes use of a simple linear filter to eliminate the effects of periodic variation A new series Z is produced whose t^{th} value is the average of Y_t and the k values of Y before and after time t, then the output series will be smoother than Y since the consecutive values of Z will have many values of Y in common in their averages. This is explained by the following formula:

$$\text{MA}(y_t) = \frac{1}{2k+1}\sum_{j=-k}^{k} y_{t+j}. \tag{3.18}$$

The symmetric MA smoother is a special case of the general idea of using linear smoothers, where new values are weighted averages of old values centred at the time point of interest. An obvious extension would be to use different weights. The weights would be greater for Y's near the time point t and smaller farther away from t.

If it is suspected that there is a linear trend and a sinusoidal cycle (seasonality) of length S in the data, a regression model (equation) with response Y_t on linear and/or sinusoidal functions of t would be used to describe the series and/or to remove trend and sinusoidal cycles.

$$Y_t = \beta_0 + \beta_1 t + \beta_2 \cos(2\pi t/S) + \beta_3 \sin(2\pi t/S) + \varepsilon_t\,, \tag{3.19}$$

where β_1 is indicates the strength of a linear trend. The strength of the sinusoid can be measured by the following quantity

$$\text{Strength}_{sinusoid} = \sqrt{\hat{\beta}_2^2 + \hat{\beta}_3^2}\ . \tag{3.20}$$

Changing variance

Diagnostic procedures such as the inspection of plots of residuals may suggest that even the best-fitting standard linear time series model is failing to provide an adequate fit to the data. A common reason for this is the non-constant variability which increases with level and may not be clearly visible in some plots. The non-stationarity of this nature is usually handled by transforming the response variable using transformation techniques such as Power and/or Box-Cox family of transformations (see e.g. Tukey, 1957; Box and Cox, 1964). In some cases we may also have the variability changing independently of the level. This problem may be handled by making use of the models which allow for non-stationary variance, e.g. Autoregressive Conditional Heteroscedasticity (ARCH) and Generalized Autoregressive Conditional Heteroscedasticity (GARCH) Models, not covered in this thesis. These models have been applied to a wide range of time series analyses, especially in modelling financial time series data (volatility of stock prices) where they are believed to be more successful in handling heteroscedasticity in the series (see e.g. Engle, 2001). In this thesis Power (Box-Cox) transformations are used.

Power Transformations

Suppose the variance of a non-stationary Y_t process changes with its level according to the following expression:

$$\mathrm{Var}(Z_t) = c \times h(\mu_t)\,, \tag{3.21}$$

where $c > 0$ is a constant and h some function.

Let $\mathrm{V}(Y_t)$ be some function which has a constant variance, and $\mathrm{V}(\mu_t)$ and $\mathrm{V}'(\mu_t)$ respectively be the value and derivative of $\mathrm{V}(Y_t)$ evaluated at μ_t. Using a first order Tailor series expansion about μ_t, we have:

$$\mathrm{V}(Y_t) \simeq \mathrm{V}(\mu_t) + \mathrm{V}'(\mu_t)(Y_t - \mu_t). \tag{3.22}$$

Thus, we have

$$\mathrm{Var}[\mathrm{V}(Y_t)] = [\mathrm{V}'(\mu_t)]^2\mathrm{Var}(Y_t) = c[\mathrm{V}'(\mu_t)]^2 h(\mu_t). \tag{3.23}$$

Then, $\mathrm{V}(Y_t)$ must be such that

$$\mathrm{V}'(\mu_t) = \frac{1}{\sqrt{h(\mu_t)}} \text{ or } \mathrm{V}(\mu_t) = \int \frac{1}{\sqrt{h(\mu_t)}} d\mu_t \tag{3.24}$$

Again if we suppose that the standard deviation is proportional to the level, i.e. $\mathrm{Var}(Y_t) = c^2{\mu_t}^2$, then

$$\mathrm{V}(\mu_t) = \int \frac{1}{\sqrt{h(\mu_t)}} d\mu_t = \ln(\mu_t), \tag{3.25}$$

in which case a logarithmic transformation will give a constant variance.

A logarithmic transformation method is commonly employed to obtain a more homogeneous variance of a univariate time series (see e.g. Graggs et al., 1999). In this way, the implications for forecasting may be quite good if the log transformed series is well described by a fitted model and the optimal forecasts for the original variable obtained. We may easily reverse the log transformation by applying the exponential function to the forecasts and thereby obtain forecast values of the original variable which is generally more efficient under ideal conditions.

If the variance of the series is proportional to the level, i.e. $\mathrm{Var}(Y_t) = c\mu_t$, then the function $\mathrm{V}(\mu_t)$ takes the following form:

$$\mathrm{V}(\mu_t) = \int \frac{1}{\sqrt{h(\mu_t)}} d\mu_t = 2(\mu_t)^{1/2}. \tag{3.26}$$

Hence, a square root transformation $\sqrt{Y_t}$ will give a constant variance in this particular case. Power transformation is a simple but often effective way to stabilize the variance of the series across time.

A reciprocal transformation may also be needed when the standard deviation is proportional to the square of the level, i.e. if $\mathrm{Var}(Y_t) = c^2{\mu_t}^4$ so that

$$\mathrm{V}(\mu_t) = \int \frac{1}{\sqrt{{\mu_t}^4}} d\mu_t = -\frac{1}{\mu_t}. \tag{3.27}$$

In general the variance can be stabilized by using the power transformation

$$\mathrm{V}(Y_t) = Y_t^{(\lambda)} = \frac{X_t^{\lambda}-1}{\lambda}. \tag{3.28}$$

The minimum value of the following preliminary sums of squares for various values of λ, can be used to suggest the appropriate transformation:

$$S(\lambda) = \sum_{t=1}^{n}(Y_t^{(\lambda)} - \hat{\mu}_t)^2, \tag{3.29}$$

where $\hat{\mu}_t$ is the sample mean of the transformed series.

This family of power transformations was introduced by Tukey (1957). The transformed values are a monotonic function of the observations over some admissible range. Such transformations may also be indexed as:

$$y_t^{(\lambda)} = \begin{cases} y_t^{\lambda} & \text{if} \quad \lambda \neq 0 \\ \log(y_t) & \text{if} \quad \lambda = 0 \end{cases} \quad y_t > 0 \tag{3.30}$$

Box-Cox Family of Transformations

The family of transformation (3.30) was modified by Box and Cox (1964) to take account of the discontinuity at $\lambda = 0$, such that

$$y_t^{(\lambda)} = \begin{cases} (y_t^{\lambda} - 1)/\lambda; & \lambda \neq 0 \\ \log y_t, & \lambda = 0 \end{cases} \quad y_t > 0 \tag{3.31}$$

For unknown λ, we have the following

$$\boldsymbol{y}^{(\boldsymbol{\lambda})} = (y_1^{(\lambda)}, y_2^{(\lambda)}, \dots, y_n^{(\lambda)})' = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \tag{3.32}$$

where $\mathbf{X}$ is a matrix of known constants, $\boldsymbol{\theta}$ is a vector of unknown parameters associated with the transformed values and $\boldsymbol{\varepsilon}\sim\mathrm{MVN}(0, \sigma_\varepsilon^2 I_n)$ is a vector of random errors. Since the transformation

(3.29) is valid only for $y_t > 0$, modifications have had to be made for negative observations. Box and Cox (1964) proposed the shifted power transformation of the following form,

$$y_t^{(\lambda)} = \begin{cases} [(y_t + \lambda_2)^{\lambda_1} - 1]/\lambda_1 & \text{if} \quad \lambda_1 \neq 0 \\ \log(y_t + \lambda_2) & \text{if} \quad \lambda_1 = 0 \end{cases} \tag{3.33}$$

where λ_1 is the transformation parameter and λ_2 is chosen such that $y_t > -\lambda_2$. The quantity λ_2 is typically chosen to be zero. Increasing λ_2 has the effect of weakening the transformation.

For $\lambda_1 < 0$, the λ_1 in the denominator of the transformation assures that $y_t^{(\lambda)}$ is an increasing function of y_t so that a plot $y_t^{(\lambda)}$ of has the same direction of trend as y_t.

Now, since $\lim_{\lambda_1 \to 0} \frac{(y_t+\lambda_2)^{\lambda_1}-1\}}{\lambda_1} = \log(y_t + \lambda_2)$, the transformation is a continuous function in λ_1.

Other versions of the transformation have been suggested by different researchers following the Box-Cox transformation (see e.g. Yeo and Johnson, 2000).

3.1.4 Box-Jenkins Methodology

Short memory models were first introduced by Box and Jenkins (1976) and until now have become the most popular models for forecasting univariate time series data. These models have originated from the Autoregressive model (AR), the Moving Average model (MA) and the combination of AR and MA.

Autoregressive (AR) Process:

The model

$$Y_t = \delta + \sum_{k=1}^{p} \phi_k \, Y_{t-k} + \varepsilon_t \tag{3.34}$$

is called the non-seasonal autoregressive model of order p, written as AR(p), where ε_t is the random shock or error assumed to be distributed as $\varepsilon_t \sim NID(0, \sigma_\varepsilon^2)$. The quantities $\phi_1, \phi_2, \dots, \phi_p$ are unknown model parameters and must be estimated from sample data.

Moving Average (MA) Process:

The model

$$Y_t = \delta + \varepsilon_t + \sum_{l=1}^{q} \theta_l \, \varepsilon_{t-l} \tag{3.35}$$

is called the non-seasonal moving average model of order q, written as MA(q). The quantities $\theta_1, \theta_2, \dots, \theta_q$ are model parameters that must be estimated from sample data. The random shock or error ε_t is again assumed to be distributed as $\varepsilon_t \sim \mathrm{NID}(0, \sigma_\varepsilon^2)$.

Autoregressive Integrated Moving Average (ARIMA) Process:

The Autoregressive Integrated Moving Average process, denoted by ARIMA(p, d, q), is given as:

$$\phi_p(L)(1-L)^d Y_t = \delta + \theta_q(L)\varepsilon_t \text{ or } \phi_p(L)\Delta^d Y_t = \delta + \theta_q(L)\varepsilon_t\,, \quad p, q > 0 \tag{3.36}$$

where

$\phi_p(L) = 1 - \phi_1 L - \phi_2 L^2 L - \cdots - \phi_p L^p$

$\theta_q(L) = 1 - \theta_1 L - \theta_2 L^2 L - \cdots - \theta_q L^q$

d is the order of differencing
$\Delta^d = (1-L)^d$ is the differencing operator
δ is a constant
ε_t is the error term, assumed to be white noise and normally distributed, i.e. $\varepsilon_t \sim \mathrm{NID}(0, \sigma_\varepsilon^2)$.

Differencing will results in a stationary process $Z_t = \Delta^d Y_t \sim$ARMA (p, q). The ARMA (p, q) process is generally represented in a lag operator notation as follows:

$$Y_t = \delta + \psi(L)\varepsilon_t, \tag{3.37}$$

with the Wold polynomial $\psi(L) = \sum_{k=0}^{\infty} \psi_k L^k \simeq \frac{\theta(L)}{\phi(L)}$, $\psi_0 = 1$, called cumulative impulse response with weights, ψ_k.

A *stationary* and *ergodic* ARMA(p, q) process has a mean equal to:

$$\mu = \frac{\delta}{1-\phi_1-\cdots-\phi_p}\,. \tag{3.38}$$

The autocovariances (γ_k), autocorrelations (ρ_k) and impulse response weights (ψ_k) of the ARMA (p, q) process satisfy the following recursive relationships,

$$\begin{aligned}\gamma_k &= \phi_1\gamma_{k-1} + \phi_2\gamma_{k-2} + \cdots + \phi_p\gamma_{k-p}\\ \rho_k &= \phi_1\rho_{k-1} + \phi_2\rho_{k-2} + \cdots + \phi_p\rho_{k-p}\\ \psi_k &= \phi_1\psi_{k-1} + \phi_2\psi_{k-2} + \cdots + \phi_p\psi_{k-p}\end{aligned} \tag{3.39}$$

for $k = q + 1, q + 2, \ldots$

When the seasonal components are included in the model, the model is called Seasonal Autoregressive Integrated Moving Average (SARIMA), written as ARIMA$(p, d, q) \times (P, D, Q)_S$. The SARIMA model reduces to a pure ARIMA(p, d, q) if there is no seasonal effect. The generalized form of the ARIMA$(p, d, q) \times (P, D, Q)_S$ model is given by:

$$\begin{aligned}&\Phi_P(L^S)\phi_p(L)(1-L^S)^D(1-L)^d y_t = \delta + \Theta_Q(L^S)\theta_q(L)\varepsilon_t \text{ or}\\ &\Phi_P(L^S)\phi_p(L)\Delta_S^D\Delta^d y_t = \delta + \Theta_Q(L^S)\theta_q(L)\varepsilon_t\,,\end{aligned} \tag{3.40}$$

$$p, q, P, Q > 0$$

where

Δ^d is the ordinary differencing operator
D is the seasonal order of differencing
Δ_S^D is the lag S seasonal differencing operator, i.e. $\Delta_S^D y_t = y_t - y_{t-S}$
$\Phi_P(L^S) = 1 - \Phi_{S,1}L^S - \Phi_{S,2}L^{2S} - \cdots - \Phi_{S,P}L^{PS}$
$\Theta_Q(L^S) = 1 - \Theta_{S,1}L^S - \Theta_{S,2}L^{2S} - \cdots - \Theta_{S,Q}L^{QS}$
ε_t is the error, assumed to be white noise and normally distributed, i.e. $\varepsilon_t \sim \text{NID}(0, \sigma_\varepsilon^2)$.

It is notable that the SARIMA model given by Equation (3.40) assumes that there is a significant parameter(s) as a result of the multiplication between non-seasonal and seasonal parameters. Such a model is called a *multiplicative* SARIMA model. The SARIMA model may also be *additive*. A Seasonal Autoregressive Integrated Moving Average (SARIMA) model is said to be additive if the non-seasonal and seasonal factors work additively, i.e. if $\left[Q_P(L^S) + \phi_p(L)\right]$ and/or $[\Theta_Q(L^S) + \theta_q(L)]$ give better results than the multiplicative cases.

Model Identification

ARIMA model identification:

The values for p and q in ARIMA(p,d,q) are identified based on the behaviour of SACF and SPACF. The SACF of an AR(p) process must dampen out and its SPACF cut off after lag p while the SACF of an MA(q) process must be willing to cut off after lag q and the SPACF dampen out. If neither the SACF nor the SPACF cuts off, then some ARMA (p, q) model will be identified. The values $p = q = 1$ are usually taken for a start.

ARIMA model identification:

i. The number of AR and MA parameters, p and q respectively, in this model is determined as explained for an ARMA process.

ii. The number of seasonal AR and MA parameters (P and Q) are determined by inspecting the sample ACF and PACF at multiples of L (i.e. $L, 2L, 3L, \dots$), the seasonal lags, as follows:

    - If the sample ACF is non-zero at lags $L, 2L, \dots, QL$ and cuts off after lag QL and the sample PACF damps out, then Q seasonal parameters are included.
    - If the sample ACF damps out and the sample PACF is non-zero at lags $L, 2L, \dots, PL$ and cuts off after lag PL, then P seasonal parameters must be included.
    - If both the sample ACF and PACF cut off (after lag QL and PL respectively), the parameters Q and P must be chosen according to which of the functions cuts off more abruptly. If the sample ACF cuts off more abruptly, then Q seasonal parameters are included. If the sample PACF cuts off more abruptly, then P seasonal parameters are included.

- If both the sample ACF and PACF damp out, we start with a model with one AR and one MA seasonal parameter and then increase the number of seasonal parameters if necessary.

Model Selection

Different models could be tentatively chosen that seem to provide statistically adequate representation of the data. The selection of the parsimonious (best) model is carried out using the *Information Criteria*, also more generally known as the *Penalty Function Criteria*. The Box-Jenkins method is characterized as being subjectively inclined or biased with an identification procedure that mainly relies on visual measures such as on the inspection of the autocorrelation plots of the data. Two penalty functions were implemented: *Akaike's Information Criterion* (AIC)and the *Schwarz's Bayesian Criterion* (SBC), (Akaike, 1983; Schwarz, 1978).

Suppose a model with m parameters is fitted to a time series. The quality of the model fitting, with respect to parsimony, can be assessed by calculating a penalized likelihood criterion

$$\mathrm{AIC}(m) = -2\ln(\text{maximum likelihood}) + 2m. \tag{3.41}$$

The logarithm of the likelihood function (L) derived from the assumption that $\varepsilon_t \sim \mathrm{IID}(0, \sigma_\varepsilon^2)$ is given by:

$$\ln L(\boldsymbol{\theta}; \sigma_\varepsilon^2) = -\frac{n}{2} ln 2\pi\sigma_\varepsilon^2 - \frac{S(\boldsymbol{\theta})}{2\sigma_\varepsilon^2}\,, \tag{3.42}$$

where $S(\boldsymbol{\theta}) = \sum_{t=1}^{n} \varepsilon_t^2(\boldsymbol{\theta}|Y_1, \dots, Y_n)$ and $\boldsymbol{\theta}$ is a vector of model parameters.

Using $\hat{\sigma}_\varepsilon^2 = \frac{S(\widehat{\boldsymbol{\theta}})}{n} \Rightarrow n\hat{\sigma}_\varepsilon^2 = S(\widehat{\boldsymbol{\theta}})$ and replacing $\boldsymbol{\theta}$ and σ_ε^2 with $\widehat{\boldsymbol{\theta}}$ and $\hat{\sigma}_\varepsilon^2$ respectively in the above equation give the following:

$$\ln L\big(\widehat{\boldsymbol{\theta}};\ \hat{\sigma}_\varepsilon^2\big) = -\frac{n}{2}(\ln 2\pi + \ln\hat{\sigma}_\varepsilon^2) - \frac{n\hat{\sigma}_\varepsilon^2}{2\hat{\sigma}_\varepsilon^2} = -\frac{n}{2}\ln\hat{\sigma}_\varepsilon^2 - \frac{n}{2}(1 + \ln 2\pi). \tag{3.43}$$

Since the second term in the above equation is a constant, it is the same for all AIC values for the candidate models and therefore can be discarded allowing the AIC criterion to be given by

$$\mathrm{AIC}(m) = n\ln\hat{\sigma}_\varepsilon^2 + 2m. \tag{3.44}$$

Now the task would be to find a value of $m = g(\boldsymbol{\theta})$ that minimizes $\mathrm{AIC}(m)$.

A disadvantage of using AIC criterion is the possible overestimation of the order of autoregression. For this reason, a Bayesian extension to this criterion, called the *Bayesian Information criterion* (BIC), was developed. This criterion is defined by:

$$\mathrm{BIC}(m) = n\ln\hat{\sigma}_\varepsilon^2 - (n-m)\ln\left(1 - \frac{m}{n}\right) + m\ln n + m\ln\left[\left(\frac{\hat{\sigma}_Y^2}{\hat{\sigma}_\varepsilon^2} - 1\right)/m\right], \tag{3.45}$$

where $\hat{\sigma}_Y^2$ is the sample series variance.

The *Schwarz's Bayesian criterion* (SBC), similar to Akaike's *Bayesian Information criterion* (BIC) is defined by

$$\mathrm{BIC}(m) = n\ln\hat{\sigma}_\varepsilon^2 + m\ln n. \tag{3.46}$$

Model Estimation

The model estimation can be carried out by using the Maximum Likelihood (ML) method to estimate the parameters. The ML function has the following general form

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} f(y_i|\theta_1, \dots, \theta_r). \tag{3.47}$$

The log-likelihood is taken to simplify derivatives when finding extreme value(s):

$$\ln L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \ln f(y_i|\theta_1, \dots, \theta_r). \tag{3.48}$$

Then, for a single parameter model, we will find the value of the parameter which maximizes the log-likelihood function, i.e. the value of θ such that:

$$\frac{d\ln(L)}{d\theta} = 0. \qquad (3.49)$$

For multiple (r) parameters, we find the values that satisfy all partial derivatives set to zero, i.e. the values of $\theta_1, \theta_2, \dots, \theta_{\mathrm{r}}$ such that:

$$\frac{\partial\ln(L)}{\partial\theta_1} = 0, \dots, \frac{\partial\ln(L)}{\partial\theta_r} = 0. \qquad (3.50)$$

Diagnostic Checking

Once a significant model has been obtained, the model is next tested for adequacy. A recommended way to check model adequacy is by examining the model residuals obtained. Commonly, if a fitted model is correct, the observed residuals $\hat{\varepsilon}_t$ should behave in much the same way as a white noise process, i.e. the following must be satisfied:

- The residuals should be independent and identically distributed normal random variables with mean 0 and finite variance σ_ε^2, i.e. $\varepsilon_t \sim \mathrm{IID}(0, \sigma_\varepsilon^2)$
- A plot of the standardized residuals $\hat{v}_t = (\hat{\varepsilon}_t - \bar{\varepsilon}_t)/\bar{\sigma}_\varepsilon$ versus t should show a random scatter (no particular pattern) about the line $\hat{v}_t = 0$ and the related normal probability plot should not violate the normality assumption. The independence assumption can be checked by analysing the sample ACF and PACF.

According to Ljung and Box (1978), in order to determine whether the first K sample autocorrelations indicate the adequacy of the model, the following hypothesis testing is used,

$H_0: r_1 = r_2 = \cdots = r_k = 0.$

The test statistics used are called the Box-Pierce statistic and the Ljung-Box statistic given respectively by

$$Q = n' \sum_{k=1}^{K} r_k^2\,(\hat{\varepsilon}) \tag{3.51}$$

and

$$Q^* = n'(n'-k) \sum_{k=1}^{K} (n'-k)^{-1} r_k^2\,(\hat{\varepsilon}) \tag{3.52}$$

In the second of the above equations, $n' = n - d$ and n is the number of observations in the series, d is the degree of non-seasonal differencing and $r_k^2(\hat{\varepsilon})$ is the sample autocorrelation of the residual at lag$k$. However, it has been theoretically proved that Q^* is the better of the two statistics and hence Q^* is recommended for testing model adequacy. Therefore, the hypothesis H_0 can be rejected for the adequacy of the model if the following holds:

$$Q^* > c_{[a]}^2(K-m), \tag{3.53}$$

where m is the number of model parameters and Q^* has an approximate Chi-Square distribution. Alternatively, H_0 can be rejected if the corresponding p-value is less than α, a pre-set significance level.

In spite of this, the goodness of fit can also be checked by simply examining the sample residual autocorrelation (SRA) and partial autocorrelation (SRPA) plots. If most of the sample autocorrelation coefficients of the residuals are within the limit $\pm 1.96/\sqrt{N}$, where N is the number of observations upon which the model is built, we can conclude that the model is adequate. In other words, if there are no spikes in the SRAC and SRPAF plots, which is an indication of a white noise distribution for residuals, then a model is a good fit to the data.

3.2 Frequency Domain Analysis

The frequency domain analysis is an alternative time series analysis approach which describes the fluctuations of time series in terms of the sinusoidal behaviour at various frequencies. This dimension of time series is concerned mostly with estimation and inference concerning the spectral density function and hence periodicities present in the data. As in the time domain approach, the frequency domain analysis requires that the series is stationary. While in the time domain analysis, functions such as autocorrelations and partial autocorrelations are used to study the evolution of a time series through parametric models, the spectral function is used in frequency domain analysis and its estimator, the periodogram, is the fundamental tool for studying periodicities in the data (see e.g. Schuter, 1987). If it is suspected that a time series contains a periodic sinusoidal component with a known wavelength, then the natural model is:

$$Y_t = A\cos(\lambda t + \delta) + \varepsilon_t, \qquad (3.54)$$

where λ (measured in radians, i.e. π radians $= 180^0$) is the frequency of the sinusoidal variation, A is the *amplitude* of the variation, δ is the *phase* and $\{\varepsilon_t\}$ denotes a white noise process. The angular frequency λ is termed the 'frequency' mainly for easier handling of mathematical formulae.

The number of cycles per unit time, referred to as *frequency* and denoted by f, is given by $f = \lambda/2\pi$ and is mainly used to interpret the results of a data process. The *period (wavelength)* is given by $1/f$ or$2\pi/\lambda$. For example, if a sinusoidal function has angular frequency $\lambda = \pi/3$, then $f = 1/6$ and the wavelength is 6.

A *periodic* function $f(t)$ is said to have a period S if for all t, $f(t+S) = f(t)$, where $S > 0$. The smallest value of S is called the period of $f(t)$. For example, $f(t) = \sin t$ has a period 2π.
The following theorem on the characterization of the autocovariance function is of particular importance.

Theorem 3.1 (*Herglotz's theorem*): A real valued sequence $\{c(k), k = 0, \pm 1, \pm 2, \dots\}$ is an autocovariance of some stationary time series $\{Y_t\}$ if and only if there exist a non-decreasing, right continuous, non-negative and bounded function $F(\lambda)$ on $\lambda \in [-\pi, \pi]$ with
$F(\lambda) = F(\pi) - F(-\lambda^-)$, such that

$$c(k) = \int_{-\pi}^{\pi} e^{i\lambda k} d\,F(\lambda). \qquad (3.55)$$

$F(\lambda)$ is called the spectral distribution function of $\{Y_t\}$.

If the function $F(\lambda)$ is differentiable such that $dF(\lambda) = f(\lambda)$, then the function $f(\lambda)$ is called the spectral density function.

Spectral Density Function and Periodogram

Restricting negative frequencies, i.e. $F(\lambda) = 0$ for $\lambda < 0$, the autocovariance function, $\gamma(k)$, is given by

$$\gamma(k) = \int_0^\pi \cos(\lambda k) dF(\lambda) = \int_0^\pi \cos(\lambda k) f(\lambda) d\lambda \,, \qquad (3.56)$$

where $\gamma(0) = \sigma_Y^2 = \int_0^\pi f(\lambda) d\lambda = F(\pi)$.

The inverse relationship of Equation (3.56) is given by

$$f(\lambda) = \frac{1}{\pi} \sum_{k=-\infty}^{\infty} \gamma(k) e^{i\lambda k} = \frac{1}{\pi} [\gamma(0) + \sum_{k=1}^{\infty} \gamma(k) \cos(\lambda k)] \,, \qquad (3.57)$$

i.e. the Fourier transform of the auto-covariance function.

Spectrum Estimation

The estimation of the spectrum of the series is rooted in Fourier analysis and making use of the Fast Fourier Transform (FFT), from which its estimator (periodogram) has been developed.
The Fourier series of a function S_t is given by

$$S_t = \frac{1}{2} a_0 + \sum_{n=1}^{\infty} a_n \cos(nt) + \sum_{n=1}^{\infty} b_n \sin(nt) \qquad (3.58)$$

where

$$a_0 = \frac{1}{N} \sum_{k=1}^{2N} S_t$$

$$a_n = \frac{1}{N}\sum_{k=1}^{2N} S_t \cos\left(\frac{\pi n k}{N t}\right), \qquad n = 0, 1, 2, \dots$$

$$b_n = \frac{1}{N}\sum_{k=1}^{2N} S_t \sin\left(\frac{\pi n k}{N t}\right), \qquad n = 1, 2, 3, \dots$$

If $\{Y_0, Y_1, \dots, Y_{n-1}\}$ is a partial realization of a time series $\{Y_t\}$, then the Fast Fourier Transform (FFT) of $\{Y_t\}_{t=0}^{n-1}$ is defined by

$$\tilde{Y}(\lambda_k) = \sum_{t=0}^{n-1} Y_t e^{-i\lambda_k t} \tag{3.59}$$

where $\lambda_k = \frac{2\pi k}{n}$, $k = 0, 1, \dots, n-1$, are the Fourier (Harmonic) frequencies.

The inverse FFT is defined by

$$Y_t = \frac{1}{n}\sum_{k=0}^{n-1} \tilde{Y}(\lambda_k) e^{-i\lambda_k t}, \quad t = 0, 1, \dots, n-1. \tag{3.60}$$

Then the periodogram of $\{Y_0, Y_1, \dots, Y_{n-1}\}$, denoted by $\{I(\lambda_k)\}_{k=0}^{n-1}$, is given as

$$I(\lambda_k) = \frac{1}{2\pi n}\left|\sum_{t=0}^{n-1} Y_t e^{-i\lambda_k t}\right|^2 = \frac{1}{2\pi n}\left|\tilde{Y}(\lambda_k)\right|^2. \tag{3.61}$$

The periodogram (estimator of the spectral density function) is graphically displayed by a plot of $I(\lambda_k)$ against λ_k or k. The following are the properties of the periodogram, viz.

- It is asymptotically unbiased, i.e. $\lim_{n\to\infty} E[I(\lambda)] \to f(\lambda)$.
- It is an inconsistent estimator of the spectrum, i.e. $\lim_{n\to\infty} \mathrm{Var}[I(\lambda)] \nrightarrow 0$.

3.3 Time Series Harmonically Coupled SARIMA Model

The frequency domain methods of spectral analysis discussed in Chapter 3 are based on an extension of the methods of Fourier analysis (Harmonic analysis) which originate in the idea that, over a finite interval, any analytic function can be approximated, to whatever degree of accuracy is desired, by taking a weighted sum of sine and cosine functions of harmonically increasing frequencies $\lambda_j = \frac{2\pi j}{s}$. Therefore, a time series model of sine and cosine functions can be used to approximate a time series data with periodic sinusoidal behaviour.

The sinusoidal models are rooted in Fourier's theorem, which states that any periodic function can be modelled as a sum of sinusoids at various amplitudes and harmonic (Fourier) frequencies. Cycles of a regular nature are often encountered in the movements of scientific objects, where their projections could be described as simple harmonic motion with parameters A (amplitude), λ (frequency) and δ (phase displacement) as observed in model given by Equation (3.54). According to Pollock et al., (1999), astronomers were the first to apply methods of Fourier analysis to time series, and their endeavour was to detect hidden periodicities within astronomical data. Typical attempts in their study were to uncover periodicities within the activities recorded by Wolfer sunspot index and in the indices of luminosity of variable stars. However, in this thesis our aim has been to apply the same method to uncover the periodicities within solar irradiance time series data. In practice, the sinusoidal model describing the function Y_t is not usually in a simplified form as in Equation (3.54). The generalized form (given as a sum of sinusoidal components) is written as follows:

$$Y_t = \sum_{j=1}^{k} A_j \cos\left(\lambda_j t + \delta_j\right) + \varepsilon_t. \qquad (3.62)$$

The frequency is a measure in radians per unit period. The quantity $2\pi/\lambda$ measures the period of the cycle. The phase displacement, also measured in radians, indicates the extent to which the cosine function has been displaced by a shift along the time axis. Thus, instead of the peak of the function occurring at time $t = 0$, as it would with an ordinary cosine function, it now occurs at time $t = \delta/\lambda$. Therefore, an underlying cyclical component from a data sequence at time t can be described by a model of the form:

$$Y_t = \alpha \cos(\lambda t) + \beta \sin(\lambda t) + \varepsilon_t\,,\ \ t = 0, \ldots, T-1. \tag{3.63}$$

If we have $S = 2n$ number of observations per day or any other cyclical variation, then Equation (3.62) can be generalized for seasonal fluctuations (of a more complicated nature) comprising the full set of harmonically related frequencies to take the form:

$$Y_t = \sum_{j=0}^{n}\{\alpha_j \cos(\lambda_j t) + \beta_j \sin(\lambda_j t)\} + \varepsilon_t\,, \tag{3.64}$$

with the harmonic scale $\lambda_j = 2\pi j/S, j = 0, \ldots, n = S/2$ in the interval $[0, \pi]$, and a_t is a residual element or white noise in the underlying process. The angular velocity $\lambda_j = 2\pi j/S$ relates to a pair of trigonometrical components which accomplish j cycles in the S periods spanned by the data. The highest harmonic frequency $\lambda_n = \pi$ corresponds to the so-called Nyquist frequency. The presence of regular harmonic components in a data series can be detected by estimating the periodogram. If in a periodogram analysis a particular intensity $I(\lambda_j)$ is the largest one, we can test the hypothesis whether the parameters α and β are indeed zero at this frequency, i.e.

H_0: $\alpha = \beta = 0$ ($\{Y_t\}_1^n$ is white noise)
H_1: $\alpha \neq 0$ and/or $\beta \neq 0$ ($\{Y_t\}_1^n$ contains a periodic component)

The above pair of hypotheses makes use of the Fisher's Kappa statistic (see e.g. Davis, 1941). The distribution of this statistic was derived by Fisher as a ratio of the largest periodogram ordinate divided by the mean of all 2 degrees of freedom ordinates. The test decision is made using the critical values for the Fisher's Kappa statistic (see e.g. Fuller, 1976). While Fisher's Kappa statistic tests the significance of the single largest periodogram ordinate, the Bartlett's Kolmogorov-Smirnov statistic generally tests for multiplicities of periodicities (Bartlett, 1963). Hence, it detects some more general departures from white noise. The usual F-test can also be used to test the significance of any periodogram ordinate of interest, e.g. the second largest ordinate, $I(\lambda_j)$. A practical example of this can be found in Chapter 6, section 6.6. It can also be shown that $U = (I(\lambda_h)\pi)/\sigma^2 \sim \chi_2^2$, for $h \neq n/2$ implying that $I(\lambda)/f(\lambda) \sim \chi_2^2$ for a white noise process with $f(\lambda) = \sigma^2/\pi$ although this result generalizes to spectra that are non-constant. Under the null hypothesis

$U = I\left(\lambda_{\mathrm{g}}\right)/\sigma_{\varepsilon}^2 \sim \chi_2^2$ for all $\mathrm{g} = 1, \dots, \left[\frac{n}{2}\right]$ independently of $V_h = \frac{S^* - I(\lambda_h)}{\sigma_{\varepsilon}^2} \sim \chi_{n-3}^2$, where $S^* = \sum_{k=1}^{[n/2]} I(\lambda_h)$. Consequently the test statistic

$$F = \frac{(U/2)}{(V_h/(n-3))} = \left(\frac{n-3}{2}\right)\left(\frac{I(\lambda_h)}{S^* - I(\lambda_h)}\right) \sim F_{2,n-3}. \tag{3.65}$$

The deterministic component and SARIMA model components are combined to construct harmonically coupled SARIMA (HCSARIMA) models to model and forecast the resulting mixture of stochastic and deterministic components of solar radiation recorded at the earth's surface. The sinusoidal function is evaluated at determined harmonic frequencies (λ_i) and hence the name *"Harmonically Coupled SARIMA Models"* or simply "HCSARIMA". Thus, the generalized form of HCSARIMA can be specified as

$$Y_t = \alpha_j \cos\left(\lambda_j t\right) + \beta_j \sin\left(\lambda_j t\right) + \Phi_P(L^S)\phi_p(L)(1 - L^S)^D(1 - L)^d \varepsilon_t \, . \tag{3.66}$$

We therefore compare the two classes of models viz., SARIMA versus HCSARIMA in modelling and forecasting the horizontal solar irradiance data series under examination in this study. This approach is applied to the irradiance time series data recorded at UKZN Howard College Radiometric Station and the fitted models are then used to forecast the irradiance in the short term. To our knowledge, this approach also has never been used to model solar irradiance time series data from this particular station nor data from any station in KwaZulu-Natal, South Africa. The results of the application of these two classes of models are presented in Chapter 6. The HCSARIMA model equation is generally a composition of irradiance variable, sinusoidal variables, significant trend (T) parameters, Box-Jenkins S/ARIMA parameters as well as sinusoidal parameters.

Chapter Summary

Chapter 3 gives a detailed discussion of two main approaches to analysing time series data. These are *time domain* and *frequency domain* techniques. The first approach (time domain) makes use of the general Box-Jenkins techniques in building a model. The models developed by this approach are generally referred to as short memory processes. The stationarity of the series is achieved by integer differencing such as regular and seasonal first-order differencing as well as variance stabilising techniques such as Box-Cox family of transformations. The model development process involves the following stages and associated methods: Model identification by visual inspection of ACF and PACF plots, model estimation by maximum likelihood method, model diagnostic (residual) analysis by Box-Pierce (Ljung-Box) tests. The latter approach (frequency domain) is appropriate when fluctuations of sinusoidal nature are inherent in the series. Spectral (periodogram) analysis of the series is then carried out to search for periodicities within the data. The techniques of this dimension of time series analysis are used to develop models of sine and cosine functions for time series with sinusoidal behaviour. In concluding the chapter, we have suggested a generic model developed from combining the sinusoidal model component and the SARIMA model to construct harmonically coupled SARIMA (HCSARIMA) models. This is useful for describing a mixture of stochastic and deterministic components in the solar irradiance time series data recorded at the earth’s surface.

Chapter 4

Long Memory Processes

In the last couple of decades, long-memory processes have evolved into a vital and important part of time series analysis. The autoregressive fractionally integrated moving average (ARFIMA) process is a class of long-memory time series models that generalizes ARIMA models by allowing non-integer (fractional) values of the differencing parameter and are useful in modelling time series with long memory property (see e.g. Granger and Joyeux, 1980; Hosking, 1981). In this chapter, we present the techniques useful for the successful handling of long-range dependent data series (see e.g. Javier et al., 2012).

Long-range dependency (LRD) is a phenomenon that may arise in the analysis of spatial or time series data. It relates to the rate of decay of statistical dependence, with the implication that this decays more slowly than an exponential decay, typically a power-like decay. That is, in a long memory process, the autocorrelation of a variable decays very slowly. In other words, the autocorrelation function of a long memory process typically decays at a hyperbolic rate (Haslett and Raftery, 1989), i.e. such processes have autocovariances that are not absolutely summable (Hurst, 1951).

4.1 Short Memory and Long Memory Properties

One common way of characterizing either a short-range or long-range dependent process is in terms of their autocovariance functions. In short-range dependent processes, the coupling between values at different times decreases rapidly as the time difference increases. Either the autocovariance diminishes to zero after a certain time-lag, or it eventually decays in an exponential sense. In long-range processes there is much stronger coupling and the decay of the autocovariance is power-like and so decays slower than exponentially.

A second way of characterizing short-range and long-range dependence is in terms of the properties of sums of consecutive values and, in particular, how the properties change as the number of terms in the summation increases. In long-range dependent processes the variance and range of the run-sums are larger and increase more rapidly, compared to properties of the marginal distribution, than for short-range dependent processes. One way of examining this behaviour is making use of the rescaled range.

A long memory property is mathematically described according to the following statement: A stationary time series $\{Y_t\}$ with auto-covariance function $\{\gamma(k)\}$ is said to have long memory if $\sum_{k=0}^{\infty}|\gamma(k)| = \infty$, i.e. the sequence of partial sums $\sum_{k=0}^{\infty}|\gamma(k)|$ diverges or is not summable.

Therefore, the difference between short range dependence and long range dependence is "all short-range dependent processes are characterized by an autocorrelation function which decays exponentially fast whereas processes with long-range dependence will exhibit a much slower decay of the correlations, i.e. the autocorrelation functions typically obey some power law.

The main objective of ARFIMA model is to explicitly account for persistence by incorporating the long-term correlations in the data. The general ARFIMA(p, d, q) process is defined by

$$\Phi(L)(1-L)^d Y_t = \Theta(L)\varepsilon_t, \quad d \in \left(-\frac{1}{2}, \frac{1}{2}\right), \tag{4.1}$$

where $\Phi(L)$ and $\Theta(L)$ are respectively the autoregressive and moving-average operators, with no common roots and $\varepsilon_t \sim NID(0, \sigma^2)$ is a white noise process. Then we can, of course, define $U_t = (1-L)^d Y_t$, so that $\{U_t\}$ is an ARMA(p, q) process.

For $d \in (-\frac{1}{2}, \frac{1}{2})$, $(1-L)^{-d}$ then becomes the fractional differencing operator and can be expressed as a binomial expansion as follows:

$$(1-L)^{-d} = \sum_{j=0}^{\infty} \eta_j L^j, \tag{4.2}$$

where

$\eta_j = \frac{\Gamma(j+d)}{\Gamma(j+1)\Gamma(d)}$ and Γ is the usual gamma function.

For large values of j, $\eta_j \sim \frac{j^{d-1}}{\Gamma(d)}$.

Before the estimation of the long memory parameter d, we describe the time series in the frequency domain.

The auto-covariance function of a general ARFIMA(p, d, q) process in (4.1) is given by

$$\gamma(k) = \frac{1}{2\pi}\int_0^{2\pi} f_Y(\lambda)e^{-i\lambda k}d\lambda\,, \tag{4.3}$$

where $f_Y(\lambda)$ is the spectral density function of the process.

4.2 Spectral Density of Long Memory Process

The process defined by Equation (4.1) is stationary and invertible. Any stationary process is the sum of a regular process and a singular process (Wold, 1938). These two processes are orthogonal and the decomposition is unique. Thus, a stationary purely nondeterministic process may be expressed as MA (∞):

$$Y_t = \psi(L)\varepsilon_t = \sum_{j=0}^{\infty}\psi_j\varepsilon_{t-j}. \tag{4.4}$$

Again the spectral measure of the purely nondeterministic process (4.2) is absolutely a continuous function with respect to $\lambda \in [-\pi, \pi]$, where the spectral density of the process (4.1) may be expressed as

$$\begin{aligned} f_Y(\lambda) &= \frac{\sigma^2}{2\pi}\left|\psi\left(e^{-i\lambda}\right)\right|^2 \\ &= \frac{\sigma^2}{2\pi}\left|1 - e^{-i\lambda}\right|^{-2d}\frac{\left|\Theta(e^{-i\lambda})\right|^2}{\left|\Phi\left(e^{-i\lambda}\right)\right|^2} \quad [\text{using } \psi(z) = (1-z)^{-d}\,\Theta(z)/\Phi(z)] \\ &= \frac{\sigma^2}{2\pi}\left[2\sin\left(\frac{\lambda}{2}\right)\right]^{-2d}\frac{\left|\Theta(e^{-i\lambda})\right|^2}{\left|\Phi\left(e^{-i\lambda}\right)\right|^2} \\ &= \left[2\,sin\left(\frac{\lambda}{2}\right)\right]^{-2d} f_U(\lambda), \end{aligned} \tag{4.5}$$

where $\lambda \in (-\pi, \pi)$ and $f_U(\lambda) = \frac{\sigma^2}{2\pi} \frac{\left|\Theta(e^{-i\lambda})\ \right|^2}{\left|\Phi(e^{-i\lambda})\right|^2}$ is the spectral density of the process $\{U_t\}$.

For a special case of the ARFIMA process, with $\Phi(L) = \Theta(L) = 1$, the spectral density function is given by (Hosking, 1981):

$$f_Y(\lambda) = \frac{\sigma^2}{2\pi}\left[2\sin\left(\frac{\lambda}{2}\right)\right]^{-2d}. \tag{4.6}$$

Theorem 4.1: Let V_t be a stationary time series with spectrum $f_V(\lambda)$ and $W_t = \sum_j a_j V_{t+j}$ with $\sum a_j^2 < \infty$. Then the spectrum of W_t is given by

$$f_W(\lambda) = \left|\sum a_j e^{-i\lambda j}\right|^2 f_V(\lambda).$$

Now let us consider the process $\{U_t\}$ with a spectral density function $f_U(\lambda)$. Since U_t=$(1-L)^d Y_t$ is a fractionally differenced series, it is stationary.

The process $\{Y_t\}$ is given by

$$\begin{aligned} Y_t &= (1-L)^{-d} U_t \\ &= \sum_{j=0}^{\infty} \eta_j L^j U_{t-j} \\ &= \sum_{j=0}^{\infty} \frac{\Gamma(j+d)}{\Gamma(j+1)\Gamma(d)} L^j U_{t-j} \\ &= \sum_{j=0}^{\infty} b_j U_{t-j}. \end{aligned} \tag{4.7}$$

Since $L \in (0,1)$, $\lim_{j\to\infty} \eta_j \to \frac{j^{d-1}}{\Gamma(d)}$ and $d \in \left(-\frac{1}{2}, \frac{1}{2}\right)$, then $\sum b_j^2 < \infty$. Thus, from Theorem 4.1, it follows that the spectrum of $\{Y_t\}$ may also be written as follows:

$$f_Y(\lambda) = \left[b_j e^{-i\lambda j}\right]^2 f_U(\lambda). \tag{4.8}$$

4.3 Estimation of Long Memory Parameter

Estimating d using the Hurst parameter: Evidence for long memory process was first proposed by Hurst (1951) while testing the behaviour of water levels of the Nile River. Although Granger and Joyeux (1980) and Hosking (1981) further popularized his work, but it is Geweke and Porter-Hudak (1983) semi-parametric procedure based on properties in the frequency domain analysis that gave a far better estimate of the long memory parameter. Since then various researchers have improved upon this procedure, (see e.g. Reisen et al., 1993; Chen et al., 1994). A simple procedure for estimating the long range parameter d is using the Hurst parameter H. The Hurst parameter $H \in (0,1)$ is a measure of the extent of long-range dependence in a time series. A value of 0.5 indicates the absence of long-range dependence. The closer H is to 1, the greater the degree of persistence or long-range dependence. Hence, the long memory parameter d is related to the Hurst parameter $H \in (0.1)$, by $d = H - 0.5$ (Beran et al., 1994).

The Periodogram Estimator: The periodogram estimator, denoted by $\hat{d}_p$, was proposed by Geweke and Porter-Hudak's (1983), who used the periodogram function $I(\lambda)$ as an estimate of the spectral density function in Equation (4.5) In this procedure the sample periodogram is used to estimate the spectrum at those frequencies near $\lambda = 0$. A straight line regression is fitted on the logarithm of the periodogram against a deterministic regressor. The number of observations to be included in a regression procedure is generally determined by $g(n) = n^{\alpha},\ 0 < \alpha < 1$, where n is the sample size. A detailed theoretical background of Geweke and Porter-Hudak's regression procedure is given as follows:

Let us recall Equation (4.5), i.e. the spectral density function of the long memory process. Taking $\ln(\cdot)$ on both sides of Equation (4.5) results in the following:

$$\ln[f_Y(\lambda)] = \ln[f_U(\lambda)] + d\left[-2\ln\left(4\sin^2\left(\frac{\lambda}{2}\right)\right)\right].$$

Then adding $\ln[I(\lambda_j)]$ on both sides of the above equation, and adding and subtracting $\ln[f_U(0)]$ on the left hand side and rearranging results in the following:

$\ln[I(\lambda_j)] = \ln[f_U(0)] + d\left[-2\ln(4\sin^2\left(\frac{\lambda}{2}\right))\right] + \ln\left[\frac{I(\lambda_j)}{f_Y(\lambda_j)}\right] + \ln\left[\frac{f_U(\lambda_j)}{f_Y(0)}\right]$.

This is simply a linear regression of the form:

$$Y_j = \alpha + \beta x_j + \varepsilon_j,$$

where

$Y_j = \ln[I(\lambda_j)]$ is the dependent variable

$\alpha = \ln[f_U(0)]$ is the intercept

$\beta =$ the slope coefficient

$x_j = -2\ln\left[4\sin^2\left(\frac{\lambda}{2}\right)\right]$ is a deterministic regressor

$\varepsilon_j = \ln\left[\frac{I(\lambda_j)}{f_Y(\lambda_j)}\right]$ is the disturbance or error term.

The term $\ln\left[\frac{f_U(\lambda_j)}{f_Y(0)}\right]$ becomes negligible when the frequency ordinates λ_j are close to zero. The least squares estimator of d is given by

$$\hat{d}_p = \hat{\beta} = \frac{\sum_{i=1}^{g(n)}(x_i - \bar{x})y_i}{\sum_{i=1}^{g(n)}(x_i - \bar{x})^2}\,, \tag{4.9}$$

where $g(n) = \sqrt{n}$ is the number of observations (periodogram ordinates) to be included in the regression procedure.

It has been shown that $\hat{d}_p \sim N\left(d, \frac{\pi}{6\sum_{i=1}^{g(n)}(x_i - \bar{x})^2}\right)$ and $\lim_{n\to\infty} E(\hat{d}_p) = d$.

Hence, $\frac{\hat{d}_p - d}{\sqrt{var(\hat{d}_p)}} \sim N(0,1)$.

The Smoothed Periodogram Estimator: The disadvantage of Geweke and Porter-Hudak's regression procedure is making use of the periodogram that is an inconsistent estimator of the spectrum. For this reason, researchers like Reisen et al. (1993) and Chen et al. (1994) conducted a study in an attempt to achieve consistency with some degree of success by smoothing (averaging or applying lag windows), hence the name smoothed periodogram. They applied the lag windows technique to smooth the periodogram. This estimator is simply obtained by replacing the spectral density function (4.5) by the smoothed periodogram function with the Parzen lag window with truncation point $m = n^{\beta}$, $0 < \beta < 1$, and $g(n)$ is selected in the similar way explained previously.

Properties of Long Memory Process: The sign of the long memory parameter for a series with long range dependence can simply be predicted by inspection of the spectrum based on the following properties in the frequency domain analysis (see e.g. Geweke and Porter-Hudak, 1983):

- For values of $d > 0$, the ACF of ARFIMA time series decays very slowly and its spectrum typically diverges to infinity at frequency $\lambda = 0$, i.e. $\lim_{\lambda \to 0} f_Y(\lambda) = \infty$.
- For values of $d < 0$, the spectrum of the series at $\lambda = 0$ is equal to zero, i.e. $f_Y(0) = 0$.
- The spectrum of the differenced series vanishes at $\lambda = 0$. This is an indication of over-differencing.

Typical Model Building Procedure

For the use of the regression techniques the following simply steps may be followed to identify and estimate an ARFIMA(p, d, q) model for a set of time series data. If $\{Y_t\}$ is a time series defined by an ARFIMA(p, d, q) model given in Equation (4.1), then $U_t = (1 - L)^d Y_t$ is an ARMA(p, q) process and $W_t = \frac{\phi(L)}{\theta(L)} Y_t$ is an ARFIMA$(0, d, 0)$ process. A general procedure for estimating the model parameters is detailed as follows:

1. Estimate d in the ARFIMA(p, d, q) model and denote the estimate by $\hat{d}$.
2. With the estimate $\hat{d}$, $\widehat{U}_t = (1 - L)^{\hat{d}} Y_t$ is computed.
3. The use of general Box-Jenkins modelling procedure for the tentative model identification and estimation of parameters ϕ and θ in the process $\phi(L)\widehat{U}_t = \theta(L)\varepsilon_t$.
4. Computing $\widehat{W}_t = \frac{\hat{\phi}(L)}{\hat{\theta}(L)} Y_t$.
5. Estimating d in the ARFIMA$(0, d, 0)$ model $(1 - L)^{\hat{d}} \widehat{Y}_t = \varepsilon_t$. The value of $\hat{d}$ obtained in this step is now the new estimate of $\hat{d}$.
6. Repeating steps 2 to 5, until the estimates of the parameters d, ϕ and θ converge.

4.4 Seasonal Fractionally Integrated Processes (SARFIMA)

The autoregressive fractionally integrated moving average process, denoted by SARFIMA$(p,d,q)(P,D,Q)_S$, is an extension of the ARFIMA(p,d,q) model (4.1) applied to a series with seasonality of length S (see e.g. Brietzke et al., 2005).

For all $D > -1$, the seasonal differencing operator, $(1 - L^S)^D$, where $S \in \mathbb{N}$ is the seasonality, is defined by the binomial expansion as follows,

$$(1 - L^S)^D = \sum_{k\geq 0}\binom{D}{k}(-L^S)^k = 1 - DL^S - \frac{D(1-D)}{2!}L^{2S} - \cdots, \tag{4.10}$$

where

$$\binom{D}{k} = \frac{\Gamma(1+D)}{\Gamma(1+k)\Gamma(1+D-k)} \ ,$$

where $\Gamma(\cdot)$ is the gamma function.

In a particular case of the SARFIMA$(p,d,q)(P,D,Q)_S$ process, where $p = q = P = Q = 0$, the process is called seasonal fractionally integrated ARIMA model with period S, denoted by SARFIMA$(0,D,0)_S$, and this process is expressed as follows,

$$\nabla_S^D = (X_t - \mu) = \varepsilon_t \ , \ t \in \mathbb{Z}.$$

Theorem 4.2 Let $\{Y_t\}_{t\in\mathbb{Z}}$ be the SARFIMA$(0,D,0)_S$ process with mean zero and $S \in N$ as the seasonal period. Then,

(i) For D $> -\frac{1}{2}$, $\{Y_t\}_{t\in\mathbb{Z}}$ is an invertible process with infinite autoregressive representation:

$$\prod(L^S)Y_t = \sum_{k\geq 0}\pi_k Y_{t-Sk} = \varepsilon_t \ ,$$

where

$$\pi_k = \frac{-D(1-D)\ldots(k-D-1)}{k!} = \frac{(k-D-1)!}{k!(-D-1)!} = \frac{\Gamma(k-D)}{\Gamma(-D)} .$$

When $k \to \infty, \ \pi_k \sim \frac{k^{-D-1}}{\Gamma(-D)}$.

(ii) For D < $\frac{1}{2}$, $\{Y_t\}_{t\in\mathbb{Z}}$ is a stationary process with an infinite moving average representation

$$Y_t = \Psi(L^S)\varepsilon_t = \sum_{k\geq 0}\psi_k\varepsilon_{t-Sk} \, ,$$

where

$$\psi_k = \frac{D(1+D)\ldots(k+D-1)}{k!} = \frac{(k+D-1)!}{k!(D-1)!} = \frac{\Gamma(k+D)}{\Gamma(D)\Gamma(k+1)}.$$

When $k \to \infty$, $\psi_k \sim \frac{k^{-D-1}}{\Gamma(-D)}$.

(iii) Assuming that $D \in \left(-\frac{1}{2}, \frac{1}{2}\right)$, the process $\{Y_t\}_{t\in\mathbb{Z}}$ has spectral density function given by

$$f_Y(\lambda) = \frac{\sigma_\varepsilon^2}{2\pi}\left[2\sin\left(\frac{S\lambda}{2}\right)\right]^{-2D}, \quad \lambda \in (0,\pi].$$

At the seasonal frequencies, for $v = 0,1,\ldots,[S/2]$, where $[x]$, means the integer part of x, it behaves as

$$f_Y\left(\frac{2\pi v}{S} + \lambda\right) \sim f_\varepsilon\left(\frac{2\pi v}{S}\right) S\lambda^{-2D}, \text{ as } \lambda \to 0.$$

(iv) The process $\{Y_t\}_{t\in\mathbb{Z}}$ has autocovariance and autocorrelation functions of order k, $k \in \mathbb{Z}$, given respectively by

$$\gamma_Y(Sk+\xi) = \begin{cases} \frac{(-1)^k\Gamma(1-2D)}{\Gamma(k-D+1)\Gamma(1-k-D)}\sigma_\varepsilon^2 = \gamma_Y(k), \; if \; \xi = 0 \\ 0, \qquad \text{if } \xi \in A \end{cases}$$

and

$$\rho_Y(Sk+\xi) = \begin{cases} \frac{\Gamma(1-D)\Gamma(k+D)}{\Gamma(D)\Gamma(k-D+1)} = \rho_Y(k), \text{ if } \xi = 0 \\ \quad 0, \qquad \text{if } \xi \in A \end{cases}$$

As $k \to \infty$, $\rho_Y(Sk) \sim \frac{\Gamma(1-D)}{\Gamma(D)} k^{2D-1}$.

(v) The process $\{Y_t\}_{t\in\mathbb{Z}}$ has partial autocorrelation function given by

$$\phi_Y(Sk+\xi, Sl+\eta) = \begin{cases} -\binom{k}{l}\frac{\Gamma(l-D)\Gamma(k-l+1-D)}{\Gamma(-D)\Gamma(k-D+1)} = \phi_Y(k,l), & \text{if}\, \eta = 0 \\ 0, & \text{if}\ \ \eta \in A \end{cases}$$

for any $k, l \in \mathbb{Z}_{\geq}$ and $\xi \in A \cup \{0\}$.

From the above expression, when $k = l$, the partial autocorrelation function of order k is given by

$$\phi_Y(Sk, Sk) = \frac{D}{k-D} = \phi_Y(k,k), \forall k \in \mathbb{Z}_{\geq}.$$

Chapter Summary

In Chapter 4, a detailed discussion of the long memory (long range dependence) property inherent in high frequency time series data is given. This is characterized by autocorrelations that decay very slowly or fail to decay at earlier lags, making it difficult to identify the suitable model from the general S/ARIMA class. The integer differencing, if used, has a drawback that it may often lead to over-differencing. For this reason, a special class of models viz., Autoregressive Fractionally Integrated Moving Average (ARFIMA) models, has been proposed in an effort to address this situation. The ARFIMA process allows non-integer (fractional) values of the differencing parameter $d \in (-0.5, 0.5)$, called long memory parameter. The simplest method of estimating the long memory parameter d is making use of the Hurst parameter $H \in (0,1)$. The Hurst parameter is a measure of the extent of long-range dependence in a time series. A value of 0.5 indicates the absence of long-range dependence. The closer H is to 1, the greater the degree of persistence or long-range dependence. Hence the long memory parameter d is related to the Hurst parameter H through the equation $d = H - 0.5$ (Beran et al., 1994). A more sophisticated technique is making use of the estimate of the spectral density function, the periodogram $I(\lambda)$. This procedure makes use of the sample periodogram to estimate the spectrum at those frequencies near zero, i.e. $\lambda_i \approx 0$. A straight line regression is fitted on the logarithm of the periodogram against a deterministic regressor. The number of observations to be included in a regression procedure is generally determined by $g(n) = n^{\alpha}$, $0 < \alpha < 1$, where n is the sample size. The properties of time series data with long memory property are: if $d > 0$, the ACF of the time series will decay very slowly and the spectrum typically diverge to infinity at $\lambda = 0$. For $d < 0$, the spectrum of the series at $\lambda = 0$ is equal to zero, i.e. $f_Y(0) = 0$. A disadvantage of using a general integer differencing is over-differencing. Over-differencing is characterized by a spectrum which vanishes at $\lambda = 0$. Just as in the short memory class we have the SARIMA model which accounts for seasonality, we also have SARFIMA process which accounts for seasonality in the long memory class.

Chapter 5

Forecasting

In time series data analysis, prediction specifically refers to the interpolation of the in-sample series values using the fitted model estimated from the sample data whereas forecasting involves making projections about the unknown future behaviour of the time series values on the basis of the observed historical performance. This is carried out by generating forecasts for the future values of the series through extrapolating trends and patterns in the past values or by extrapolating the past effect of other variables on the series. In the scientific field, the underlying data generating the system's features are effectively handled by the sophisticated time series modelling techniques and tools, used interactively to develop forecasting models customized to predict the time series with a high degree of accuracy. The principal purpose in modelling time series data is to build a model which best explains the underlying data generating process and allows the extrapolation into the future values of the time series variable under investigation. In this study, the focus is largely on analysing and generating short term forecasts for various solar irradiance time series data using the HCSARIMA model developed in Chapter 3. A comparative analysis is done with SARIMA models. As such, we make use of the various statistical techniques to help us choose between the candidate models.

Suppose that we have an observed time series $Y_1, Y_2, \dots, Y_t$ up to time t and the T-step ahead future values $Y_{t+1}, Y_{t+2}, \dots, Y_{t+T}$ are to be forecasted with a particular forecasting method. Even if the time series actually follows some assumed model, the future value of the noise is unknown. Therefore, with a correct forecasting method or model the forecast for each of the future values $Y_{t+1}, Y_{t+2}, \dots, Y_{t+T}$ is expressed as follows:

$$\hat{Y}_t(\tau) = E(Y_{t+\tau} | Y_t, Y_{t-1}, \dots), \quad \tau = 1, 2, \dots, T. \qquad (5.1)$$

The forecast value of Y_{t+k} can also be expressed as a function of the model sample parameter estimates, obtained using the sample time series data for times up to t, as follows:

$$\hat{Y}_t(\tau) = E(Y_{t+\tau}) = f\left(\hat{\boldsymbol{\theta}}, t+\tau\right), \tag{5.2}$$

where $\hat{\boldsymbol{\theta}}$ is the vector of the estimated model parameters.

In sections to follow, we discuss various time series smoothing and forecasting methods that are commonly used.

5.1 Exponential Smoothing

Exponential smoothing was first proposed by Brown (1956) and then expanded by Holt (1957). This is another common forecasting scheme to produce a smoothed time series. Exponential smoothing assigns *exponentially decreasing weights* to the older observations. In other words, *recent observations are given relatively more weight in forecasting than the older observations*, whereas in the case of moving averages, the weights assigned to the observations are the same. In exponential smoothing, however, there are one or more *smoothing parameters* to be determined (or estimated) and these choices determine the weights assigned to the observations.

Simple Exponential Smoothing (SES) is the most widely used method of all forecasting techniques. It is used for short-range forecasting, usually just one period into the future. This method also requires that the time series data pattern is approximately horizontal (i.e. there is no neither cyclic variation nor pronounced trend in the historical data). That is, the method is based on the assumption that the data fluctuates around a reasonably stable mean and is described by the model given in Equation (5.3).

If $\{Y_t\}$ represents the raw time series data and $\{S_t\}$ the output of the exponential smoothing algorithm, then the simplest form of the exponential filter with a smoothing factor $\alpha \in [0, 1]$ which creates the series $\{S_t\}$, is given by the following formulae:

$$\begin{aligned} &S_0 = Y_0 \\ &S_t = \alpha Y_{t-1} + (1-\alpha) S_{t-1},\ t > 0 \end{aligned} \tag{5.3}$$

where $S_0 = Y_0$ is the initialization of the output series. The initial value of S_t plays an important role in computing all the subsequent values. Setting it to Y_0 is one general method of initialization. According to Kalekar et al. (2004), another possibility would be to average the first four or five observations. The simple exponential smoothing formula is the adaptive forecast-updating form of the exponential smoother. This implies that

$$S_t = \alpha \sum_{k=0}^{T-1}(1-\alpha)^K Y_{T-k} + (1-\alpha)^T S_0 \text{ for } t = 1,2,\dots,T. \tag{5.4}$$

In effect, each smoothed value is the weighted average of the previous observations, where the weights decrease exponentially depending on the value of the smoothing parameterα. The choice of the smoothing constant α determines how quickly the smoothed series or forecast will adjust to changes in the mean of the unfiltered series. For small values of α, the response will be slow because more weight is placed on the previous estimate of the mean of the unfiltered series, whereas larger values of α will put more emphasis on the most recently observed value of the unfiltered series. It is also noted that if $\alpha = 1$, then the previous observations are ignored entirely and if $\alpha = 0$, then the current observation is ignored entirely, and the smoothed value consists entirely of the previous smoothed value (which in turn is computed from the smoothed observation before it, and so on; thus all smoothed values will be equal to the initial smoothed value S_0). The in-between values of α will produce intermediate results. As an example to demonstrate the applicability of the simple exponential smoothing method, let us consider the sample series of 3rd of Feb 2010 where there are missing readings at the time points 10:11AM, 10:12AM and 10:15AM. Employing SES with a smoothing factor of $\alpha = 0.3$, we can initialize the output vector at 10:09AM with S_0=1091.644. Then the predicted value for 10:11AM is 1075.233 and 1005.152 for 10:15AM.

Other exponential smoothing methods which are not part of this thesis include double exponential smoothing for a series with trend, triple exponential smoothing for a series exhibiting both trend and seasonality, multiplicative seasonal model for a time series exhibiting multiplicative seasonality, additive model for a time series with the gradually increasing trend and a more or less constant seasonality.

5.2 Prediction Accuracy Analysis

5.2.1 Model predicted versus actual values

Accuracy Measures: The common indicators of the prediction error are Mean Bias Error (MBE), Mean Percentage Error (MPE), Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE). The Mean Biased Error (MBE) provides information on the long-term performance, over or under estimation of the model in the long run. The Mean Percentage Error (MPE) indicates the average ratio of deviations to the actual values and the Mean Absolute Percentage Error (MAPE) simply takes the absolute value of the MPE. The Root Mean Square Error (RMSE) is one of the most commonly used indicators. However, a clear disadvantage of RMSE is that, it may read a high value even if only a single measurement has high deviation from its model generated counterpart. For n error observations used to compute the mean, these indicators are defined as follows:

$$\begin{aligned}
&\text{MBE} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_i - Y_i\right)\\
&\text{MPE} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\hat{Y}_i - Y_i}{Y_i}\right)\\
&\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{Y}_i - Y_i}{Y_i}\right|\\
&\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_i - Y_i\right)^2}\ .
\end{aligned} \tag{5.5}$$

Coefficient of Determination: In statistics, the coefficient of determination, denoted by $R^2 \in [0,1]$ and pronounced R-squared, indicates how well data points fit a line or curve. It is a statistical indicator in the context of statistical models whose main purpose is to provide a measure of how well the observed outcomes are replicated by the model as the proportion of the total variation of outcomes are explained by the model. For example, an R-squared value of 0.75 would mean that the fitted model accounts for only 75% of the variability in the data. In general terms, R^2 can also be defined as a statistical measure for the goodness-of-fit. The value of R^2 is computed by

$$R^2 = 1 - \frac{SSE}{SST}, \qquad (5.6)$$

where SSE and SST are respectively the error sum of squares and the total corrected sum of squares. The error sum of squares $SSE = \sum_t (y_t - \hat{y}_t)^2$ measures the deviations of observations from their predicted values $\hat{y}_t$. The total corrected sum of squares $SST = \sum_t (y_t - \bar{y})^2$ measures the deviations of the observations from their mean $\bar{y}$. In general the higher the value of R^2, the more useful the model is. In this thesis, R-squared value is used to indicate how well the actual (in-sample) time series values (y_t) are explained by the model interpolated values ($\hat{y}_t$).

5.2.2 Forecast Error Distribution

In Section 5.2.1 we have only been concerned with making estimates for future values of the time series variable. In this section, we present methods for measuring the forecast error accuracy and estimating a confidence interval around a forecast. One obvious desirable characteristic of the forecast $\hat{Y}_t(\tau)$ is that it is *unbiased*. For an estimate to be unbiased, it must satisfy the following:

$$E[\hat{Y}_t(\tau)] = Y_t(\tau), \qquad (5.7)$$

i.e. the expected value of the forecast must be equal to expected value of the time series.

The assessment of prediction errors is always at the centre of the forecasting method evaluation and a good forecasting method is obviously the one which minimizes the distances between the predicted (forecasted) values and the actual values of the series. The error in forecasting $Y_{t+\tau}$ is mathematically expressed as:

$$e_\tau = Y_{t+\tau} - \hat{Y}_t(\tau), \qquad (5.8)$$

i.e. the difference between the estimated value and the actual value. The error e_τ is randomly distributed and its probability distribution is investigated by computing its mean and variance. On the basis of the assumption that the fitted model is correct, Equation (5.8) can be rewritten as

$$e_\tau = E(Y_{t+\tau}) + \varepsilon_t - \hat{Y}_t(\tau), \qquad (5.9)$$

where ε_t is white noise by assumption. Thus, an unbiased forecast implies that $E(e_\tau) = 0$. If the noise terms are uncorrelated, i.e. $\text{Cov}[\varepsilon_t \varepsilon_s] = 0$, the variance of the error must be given by:

$$\text{V}ar[e_\tau] = \text{V}ar[E(Y_{t+\tau}) - \hat{Y}_t(\tau)] + \text{V}ar[\varepsilon_{t+\tau}] = \sigma_E^2(\tau) + \sigma_\varepsilon^2 \tag{5.10}$$

Therefore, the variance of the error in estimating the future value $Y_{t+\tau}$, is the sum of two different variances, i.e. the one that is due to the estimation of the mean, $\sigma_E^2(\tau)$, and the other is the variance of the noise, σ_ε^2. Due to the inherent inaccuracy of the statistical methods used to estimate the model parameters and the possibility that the model is not exactly correct, the variance in the estimate of the means is an increasing function of τ.

Given n sample forecast errors, $\{e_i\}_{i=1}^n$, the sample standard deviation of the error is given by:

$$s_e = \sqrt{\frac{\sum_{i=1}^n (e_i - \bar{e})^2}{n-p}}\ , \tag{5.11}$$

where $\bar{e}$ is the sample average error and p is the number of parameters in the model. The value of s_e^2 for a given value of n is an estimate of the error variance σ_e^2. This includes the combined effects of errors in the model and the noise. If it is assumed that the random noise comes from a normal distribution, a 95% confidence interval estimate of the forecast can be approximated by:

$$\hat{Y}_{t+\tau} \pm 1.96\ s_e\ , \tag{5.12}$$

or using Student's t-distribution with $n - p$ degrees of freedom, by

$$\hat{Y}_{t+\tau} \pm t_{n-p;\,0.975} s_e. \tag{5.13}$$

5.3 Forecasting Solar Flux

Photovoltaic power production is increasing nowadays and the power output depends on the incoming radiation and on the solar panel characteristics e.g. storage system. Therefore, accurate and reliable forecast information is essential for an efficient use, the management of the electricity grid and for solar energy trading. The two main challenges to high penetration rates of PV systems are *variability* and *uncertainty*, i.e. the fact that PV output exhibits variability at all timescales (from seconds to years) and the fact that this variability itself may be difficult to predict. Thus, both issues are addressed with trends analysis and forecasting. Solar forecasting can be done on three main horizons namely; *now-casting* (forecasting 3 to 4 hours ahead), *short-term forecasting* (up to 7 days ahead) and *long-term forecasting* (months, years… ahead).

Now-casting: generally referred to as *intra-day* forecasting, is usually related to a very high temporal resolution (i.e. a forecast every 10 or 15 minutes).

Short-term forecasting for PV output: provides forecasts up to 7 days ahead. This kind of forecast is also valuable for grid operators in order to make decisions related to future power supply or demand, as well as, for electric market operators. In this thesis we present the results of short-term forecasting up to two days ahead for global horizontal irradiance $[W/m^2]$ on hourly and ten minute scales (see e.g. GeoModel Solar), with the proposed models whose outputs were post-processed with statistical approaches based on measured data. A day ahead forecasting of solar radiation on hourly scale has also been performed (see e.g. Kobayashi et al., 2013).

Long-term forecasting for PV output: usually refers to forecasting of the annual or monthly available solar resource. This is useful for energy producers and to negotiate contracts with financial entities or utilities that distribute the generated energy. In general, such forecasting is usually done at a lower scale than any of the other two approaches.

Chapter Summary

In Chapter 5, forecasting methods have been discussed with respect to their application according to specific behaviours by time series data. Such forecasting techniques are moving average and simple exponential smoothing methods for a series with no trend, double exponential smoothing for a series with trend, triple exponential smoothing for a series exhibiting both trend and seasonality, multiplicative seasonal model for a time series exhibiting multiplicative seasonality, additive model for a time series with the gradually increasing trend and a more or less constant seasonality. A good forecasting method or model is the one that gives a minimal possible forecast error. Such a method is always preferred for generating forecast values of variable. The prediction error is commonly measured by the following statistical indicators; MBE, MPE, MAPE and RMSE. The forecast errors and their confidence intervals have been discussed. We have also looked at some methods that have been used to estimate the missing values in the time series data (see e.g. Huo et al., 2010). In concluding the chapter, we have presented an overview of forecasting solar irradiance, namely now-casting (up to a few hours ahead or intra-day), short-term forecasting (up to 7 days ahead) and long-term forecasting (months and years ahead). In the following chapter, we present all experimental time series data under investigation, together with the results obtained using statistical packages, R and SAS programs.

Chapter 6

Data and Analysis Results

6.1 Solar Measurements and Meteorological Conditions

Solar irradiance measurements for UKZN HC solar radiometric station (located at 29.9° South, 30.98° East with elevation, 151.3m) were obtained from the Greater Durban Radiometric Network (GRADRAD) database, accessible at http://www.gradrad.ukzn.ac.za.The recording started as from 1 Feb 2010. This is a local radiometric database which includes other two broadband ground stations in KwaZulu-Natal, South Africa. One is located at UKZN Westville Campus and the other at Mangosuthu University of Technology. The three stations are within a 20 km radius of each other and lie on the east coast of South Africa. The solar resource at the latter mentioned station has also been assessed by other authors (see e.g. Zawilska and Brooks, 2012).

The sampling scheme employed at these stations allows the collection of high quality global, direct and diffuse irradiance measurements with the aid of thermopile instruments and a common software format to facilitate comparison of data. Instantaneous readings were made every six-second intervals and then averaged over a one-minute period. At UKZN HC solar meteorological station, global irradiance readings which we consider in this study were made with a precision spectral pyranometer (PSP). Diffuse irradiance measurements were also made with PSP and direct irradiance with normal incident pyrheliometer (NIP).

For the data used in the study by Craggs et al. (1999), instantaneous global irradiance readings were made every minute also with pyranometer and then averaged over a ten-minute period, recorded by the datalogger. These measurements were made from the station located in the city centre of Newcastle upon Tyne, 14 km from the east coast of the UK at latitude 54859' N, longitude 1837' W and 44 m above sea level. This station has a cool temperate climate with an ambient temperature usually between -5°C and 30°C. The Durban temperatures range from 16-25°C in winter (June to August) and 23-33°C during the summer months (November to March) and so is the ambient temperature for the solar station of UKZN HC. Figure 6.1 shows a solar high-quality ground station within UKZN HC area recording solar data at 1-minute intervals.

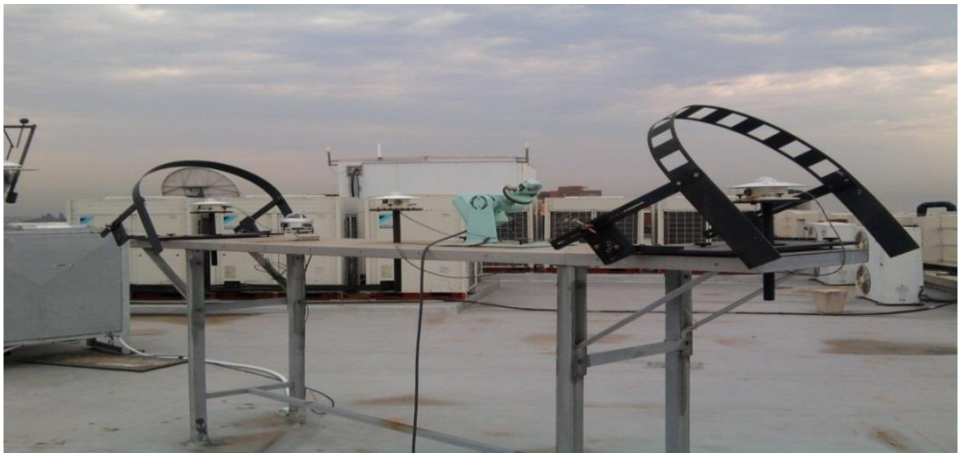

Figure 6.1: A photo of Eppley Bench solar measurement equipment at UKZN HC high-quality ground station. Location: 29.9º South, 30.98º East, Elevation: 151.3m. *Source*: Own photograph.

Shown in the photo above, is a solar tracker, PSP, with a solid shadow band for DHI only and perforated shadow band for GHI and DHI and solar tracker, NIP (green in colour), for DNI only.
In a similar study conducted by Craggs et al. (1999), global solar irradiance on the horizontal and vertical orientations for periods in two winters and two summers was examined. However, in our own study we examine global solar irradiance only on the horizontal orientation as there is no data on the vertical orientation available. At UKZN HC station, we also sampled series for the months of February and July only because at this station, February is one of the months in which we experience a summer season and July is one in which we experience winter. The degree of cloudiness at this site during these two months in 2010 and 2011 was also measured in terms of the clearness index, K_t. The classification using information in Table 2.2 was carried out and percentages of days in each category calculated, as shown in Table 6.1 below:

Table 6.1: Number of sample days, percentage of days with indicated sky conditions in terms of the median (50th percentile) clearness index, for the modelled 10-minutely and 60-minutely data sets.

Month	No. of sample days	% Clear	% Partially cloudy	% Cloudy	% Missing Data (10 min)	% Missing Data (60 min)
Feb 2010	12	17	75	8	1.36	2.32
Feb 2011	13	23	77	0	5.79	5.79
Jul 2010	13	0	69	31	0.78	2.44
Jul 2011	7	0	57	43	1.45	9.66

From Table 6.1, there is enough information to conclude that all the periods for which the modelled series were sampled, were dominated by partially cloudy sky conditions.

6.2 Data Quality

6.2.1 Missing Values

Solar irradiance measurements for UKZN HC solar radiometric station were examined over a period of 7 to 13 days for the months of February and July in the years of 2010 and 2011. These measurements were recorded on minutely time horizons and the averages of 10 minutely and 60 minutely time scales had to be obtained for modelling purposes. But the missing value problem, mainly caused by equipment failure or cleaning or equipment being offline has been encountered on a few time points. To address this problem, literature methods such as Average Nearest Observation (ANO) among others have been used for interpolating or extrapolating the missing series values within or outside a range of available data points (Gupta and Srinivasan, 2011). The ANO is the simplest method used for replacing all missing values for a given series with the mean, median, or other location statistics (e.g. percentiles) determined from the non-missing values (DeLurgio, 1998). The rest of the section presents the demonstration of the efficacy of the ANO method in estimating the missing series values and the introduction of two-directional exponential smoothing method (TES) for similar purpose and future applications.

Average nearest observation (ANO): The ANO method will replace missing values with the average of nearest previous and following observations. As an example, let us consider a series of 20 observations, from 10:10AM to 10:19AM, with missing data gaps. This series is represented as [1036.908, $missing^1$, $missing^2$, 1022.532, 964.598, $missing^3$, 1042.981, 1044.061, 1026.582, 843.026]. This is one series for which the 10-value average had to be computed for modelling purposes. For such series, the following steps in replacing the missing values were made:

[1036.908, $missing^1$ = (1036.908+1022.532)/2=1029.720, $missing^2$, 1022.532, 964.598, $missing^3$, 1042.981, 1044.061, 1026.582, 843.026]
= [1036.908, 1029.720, $missing^2$ = (1029.720+1022.532)/2=1026.126, 1022.532, 964.598, $missing^3$, 1042.981, 1044.061, 1026.582, 843.026]
= [1036.908, 1029.72, 1026.126, 1022.532, 964.598, $missing^3$ = (964.598+1042.981)/2=1003.790, 1042.981, 1044.061, 1026.582, 843.026]
= [1036.908, 1029.720, 1026.126, 1022.532, 964.598, 1003.790, 1042.981, 1044.061, 1026.582, 843.026].

It is observed that $missing^2$= 1036.908×1/4+1022.532×3/4, i.e. $missing^2$ is estimated by a weighted average of the nearest observations with higher weight given to the closer observation. It is instructive to note that this method will generate different replacement values if the time series occurs in the opposite order. For example, the reversed time series [843.026, 1026.582, 1044.061, 1042.981 $missing^1$, 964.598, 1022.532, $missing^2$, $missing^3$, 1036.908] becomes: [843.026, 1026.582, 1044.061, 1042.981, 1003.790, 964.598, 1022.532, 1029.720, 1033.314, 1036.908].

ANO method efficacy: The applicability of this method has been evaluated using the sample series of 70 minutely values from 09:00AM to 10:09AM, on this day of Feb 3, 2010 with median clearness index K_C =0.47. The gaps of various lengths of average 1.78 minutes have been created to generate a data series with missing values for the purpose of evaluating the method. The plots and results are shown below:

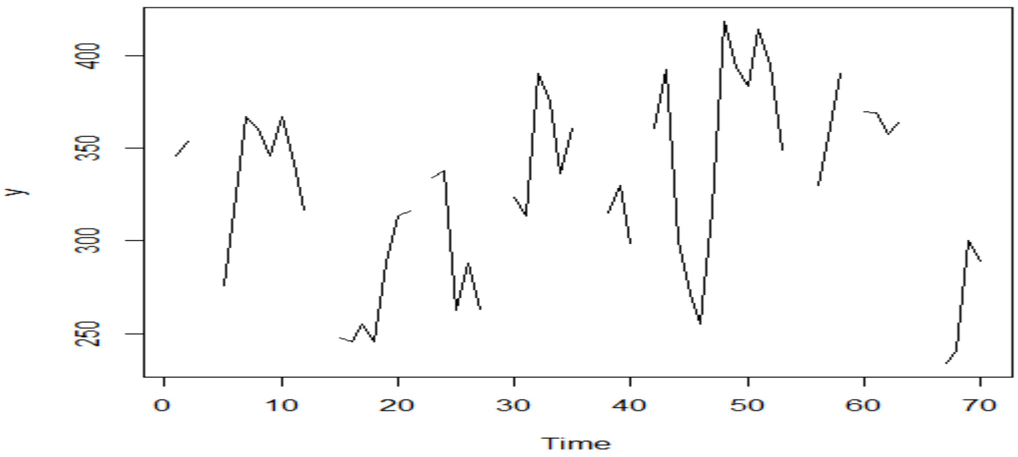

Figure 6.2: A test raw data series of global solar irradiance with data gaps (missing values) created for method testing purposes.

Table 6.2: Error evaluation in the estimation of the missing values by ANO method.

MBE	MPE	MAPE	RMSE	R-square
-8.29865	-2.03262	7.015105	26.65795	0.747023

From Table 6.2, it is apparent that the ANO method does give reasonably good results. Such statistics obviously improve with the decrease in the variability of the series values. Another good feature of this method is its ability to fill in all the values. The ANO method also serves as basis for advanced TES method discussed in the following section.

Average Nearest Observation (ANO) and Two-Directional Exponential Smoothing (TES)

The Average Nearest Observation (ANO) and Two-Directional Exponential Smoothing (TES) methods have been developed to replace routinely missing values in time series data (see e.g. Huo et al., 2010). Here, we give a detailed procedure of how the two methods work in association with each other.

The ANO method estimates the missing data point with the average of the nearest previous and the following observations. However, according to Huo et al. (2010), this method performs poorly for time series data with weak autocorrelation and/or strong daily seasonality. Where this is the case, manually entering more reasonable estimates before using the algorithm is strongly recommended. These estimates are the averages of the forward and backward exponential smoothing (ES). The procedure for obtaining the TES estimates is discussed in the next paragraph.

The TES method depends on a suitable exponential smoothing method and was developed by using Holt's linear trend algorithm method. The TES method estimates missing data points based on the autocorrelations of the time series to account for the fact that the missing values occur at non-random times. The TES method has been designed to represent both forward and backward autocorrelations in the time series. This method uses the averaged forward and backward ES estimates for predicting the missing data points and therefore can reduce the variability caused by different directions. The first step in the TES method is to generate the full set of data using the ANO method. Once the data set is generated using the ANO method, the missing values are predicted using a suitable exponential smoothing, Holt's linear trend method, in the forward and reverse direction. Therefore, the final replacement values for the missing data points are the averages of the forward and backward TES estimates, i.e. the TES method is a combination time series and is represented for missing values as:

$$TES_t = (ES_{forward,t} + ES_{backward,t})/2 \,, \tag{6.1}$$

if the value is missing. Otherwise TES_t = original value.

6.2.2 Data Comparative Methods

Before release for utility purposes, it is an essential step to perform data quality checks so as to assure that the data are within reasonable bounds. This generally requires reliable estimation algorithms for relatively accurate data generation. However, there can still be tolerances on measured data to account for possible equipment bias errors and additional sources of uncertainty in the models and algorithms themselves, according to Badescu (2008) and Gueymard et al. (2002). Below are three approaches to quality assessment of solar radiation data namely: *comparison with physical limits*, *comparison with closure* and *model comparison*.

Quality Assessment Based Upon Physical Limits: This compares measured data with estimated or defined limits to address questions such as for instance: Is the radiation component within the range of zero to the maximum possible expected value? Is the direct normal irradiance greater than zero and less than the extra-terrestrial value I_0, i.e. $0 < I_B < I_0$? Is the global horizontal irradiance I_B not greater than the vertical component of the extra-terrestrial beam? Is the diffuse irradiance more than the expected Rayleigh diffuse sky? While the possibility of one or more components can be allowed to pass such tests even when bad, however, for the most part, the physical limits tests cannot provide the level of accuracy required to assure the smooth operation of the measurement equipment, unless used with intensive human interaction.

Quality Assessment Based Upon Physical Closure: This approach makes use of the theoretical relation between the three solar components, given in Equation (2.6). This approach can be implemented directly, or more simply using irradiance values normalized to extra-terrestrial beam (I_0) values, known as clearness indices, discussed in the previous section. Equation (2.6) then takes the following form:

$$K_G = K_B + K_D. \tag{6.2}$$

The relationship between K_G and K_B at a particular site is analysed with the aid of boundaries defined by double-exponential Gompertz functions (Trouve et al., 2005). This is the family of curves, called Gompertz curves, defined by $Y = AW^{CW^{D\lambda}}$, where the choices of A, W, C and D result in proper “S” shaped boundaries around the data (Parton and Innes, 1972). Given the scatter plot of K_G versus K_B, acceptable values then fall within the analytic boundary curves. An important point to keep in mind regarding whichever approach is used, is that with the known uncertainties in measured data, a tolerance or acceptable deviation from perfect closure is needed. Typically, with measurement data uncertainties of 3% to 5% in total global and direct beam data, tolerances of $\pm 5\%$ in the balance, are generally allowed. This means tolerances of about 0.02 to 0.03 in the clearness-index approach.

Quality Assessment Based Upon Comparison with Fitted Models: This approach compares the fitted data models with clear-sky measured data to measure the magnitudes of deviations between model predicted data and actual data. The error indicators discussed in Section 5.3, are generally used for judgments in this respect.

6.3 Data analysis and plots for 60-minutely and 10-minutely averaged horizontal global irradiance

In this section, we present some of the essential details of all 60-minutely and 10-minutely averaged global (horizontal) irradiance data series, from UKZN Howard College Solar Station, relating to 7 to 13 day period in the years of 2010 and 2011. In this study, the irradiance series were examined for sample periods in two winters and two summers. Some technical details of all data series under investigation are also presented in Table 6.4. In facilitating equivalent modelling, the same cut-off times were applied to each day within the same observation period.

Table 6.3: Season, year and duration in days for the sampled series.

Season	Start date	End date	Duration
Summer 2010	01 Feb 2010	12 Feb 2010	12 days
Winter 2010	01 Jul 2010	13 Jul 2010	13 days
Summer 2011	01 Feb 2011	13 Feb 2011	13 days
Winter 2011	03 Jul 2011	09 Jul 2011	7 days

Table 6.4: Details for time series data lengths and daily cyclical lengths.

		10 min		**60 min**	
Season		**Series Length**	**Cycle length**	**Series Length**	**Cycle Length**
Summer	Feb 2010	1080	90	192	16
	Feb 2011	1092	84	182	14
Winter	Jul 2010	884	68	156	12
	Jul 2011	462	66	84	12

Presented in Figures 6.3 to 6.8 are the plots of all 60-minutely and 10-minutely averaged daylight global (horizontal) solar irradiance data series incident on the solar panels at UKZN HC radiometric station during the summer and winter seasons of 2010 and 2011. We observe that on some days, the lower levels of solar energy are experienced at this station, mainly because of overcast sky conditions.

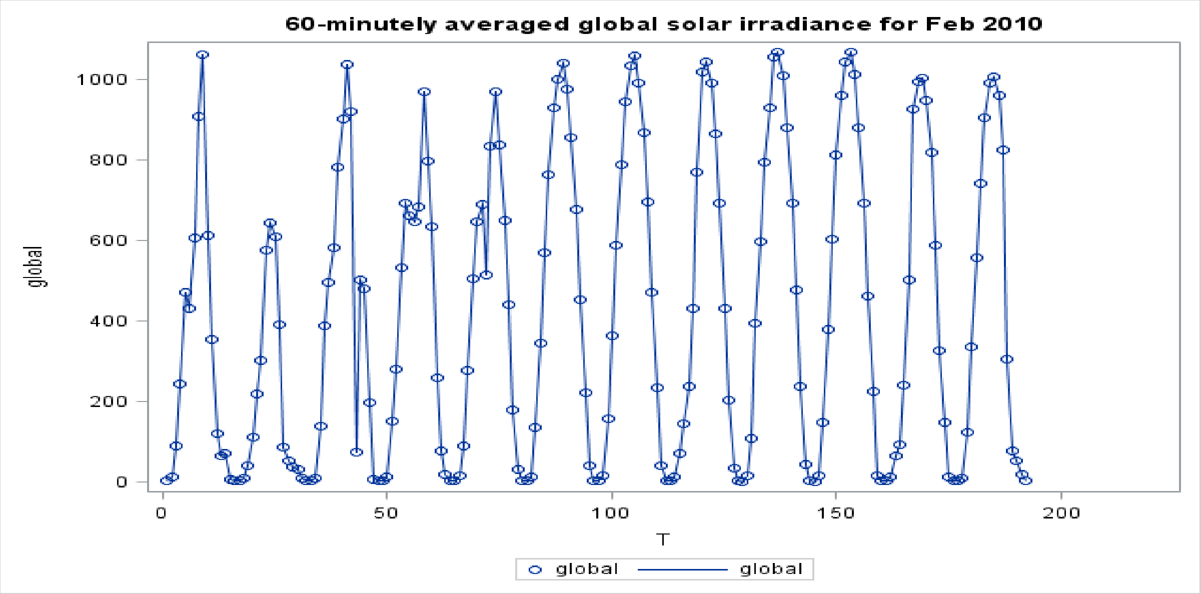

Figure 6.3: The plot of the 60-minutely averaged daylight global (horizontal) solar irradiance series over the period from 1 Feb 2010 to 12 Feb 2010.

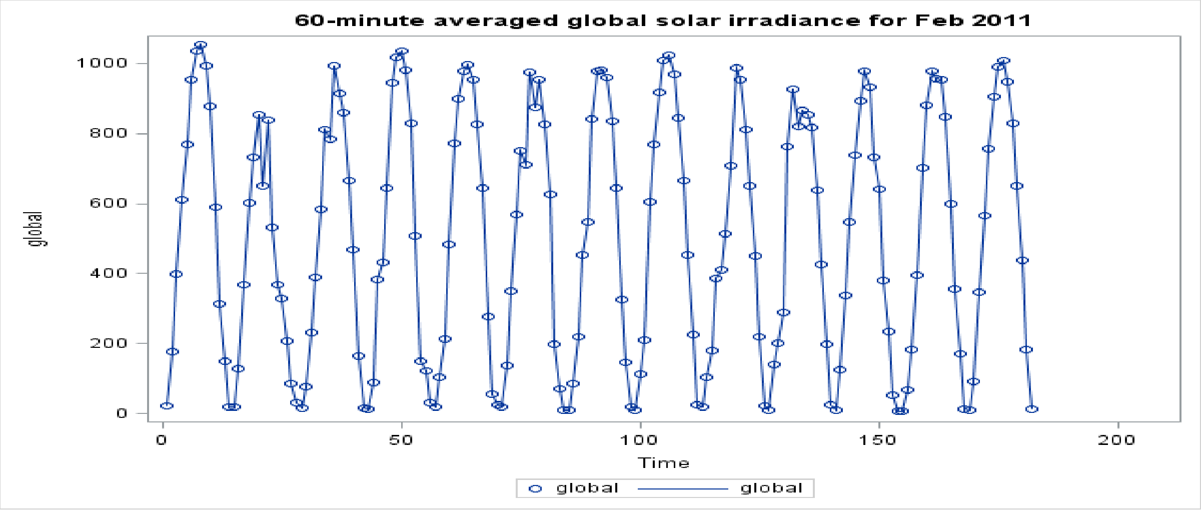

Figure 6.4: The plot of the 60-minutely averaged daylight global (horizontal) solar irradiance series over the period from 1 Feb 2011 to 13 Feb 2011.

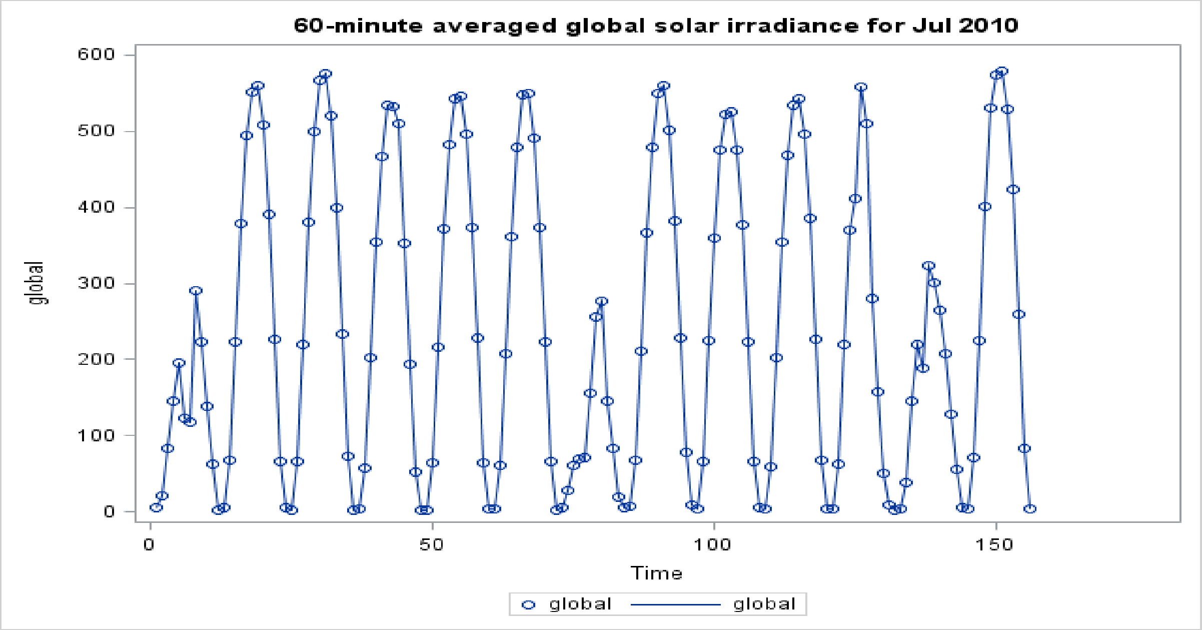

Figure 6.5: The plot of the 60-minutely averaged daylight global (horizontal) solar irradiance series over the period from 1 Jul 2010 to 13 Jul 2010.

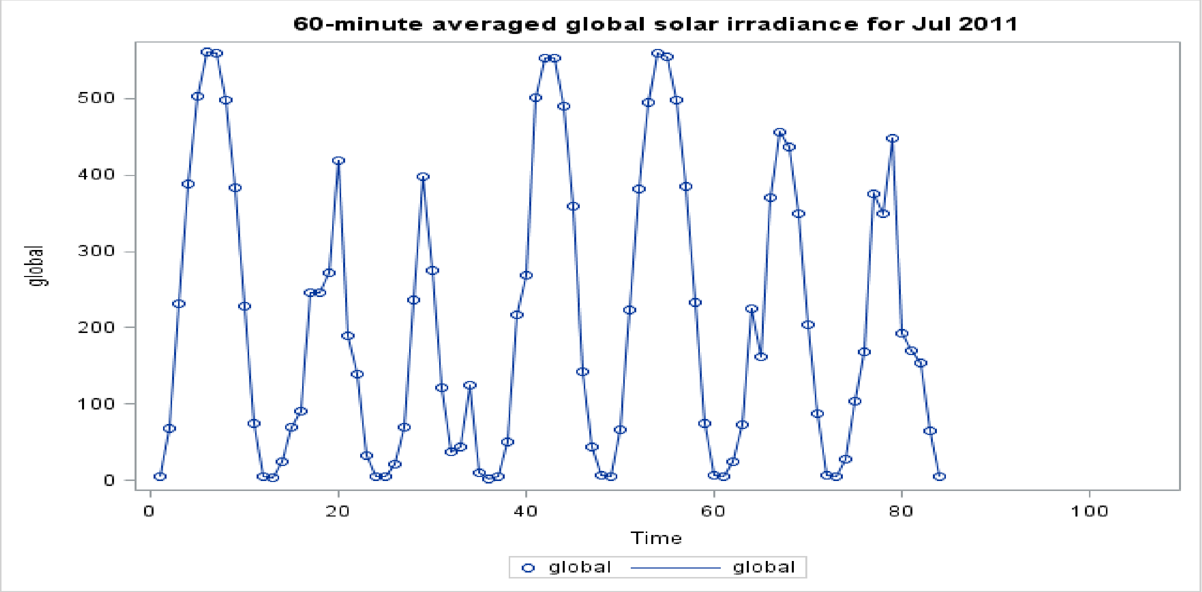

Figure 6.6: The plot of the 60-minutely averaged daylight global (horizontal) solar irradiance series over the period from 3 Jul 2011 to 9 Jul 2011.

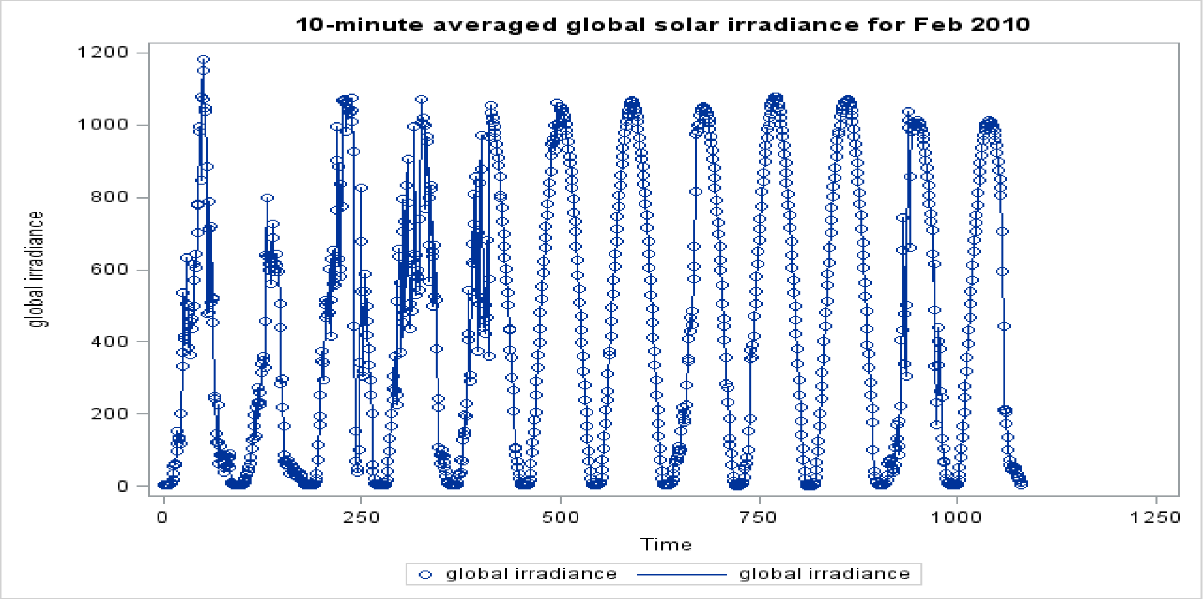

Figure 6.7: The plot of the10-minutely averaged daylight global (horizontal) solar irradiance series over the period from 1 Feb 2010 to 12 Feb 2010.

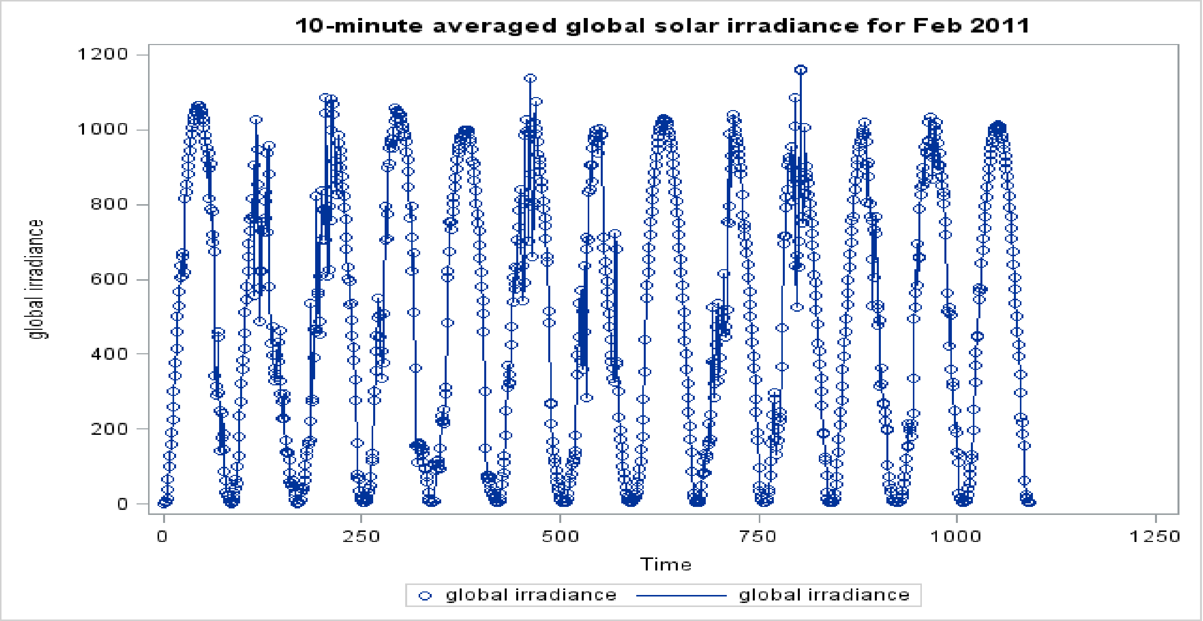

Figure 6.8: The plot of the 10-minutely averaged daylight global (horizontal) solar irradiance series over the period from 1 Feb 2011 to 13 Feb 2011.

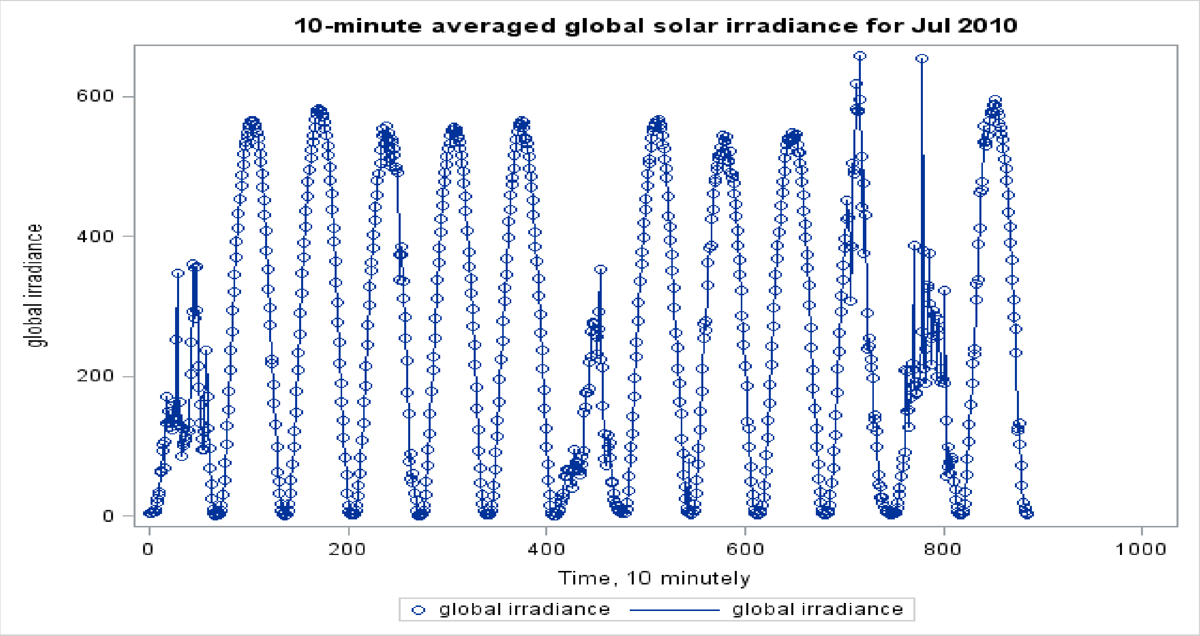

Figure 6.9: The plot of the 10-minutely averaged daylight global (horizontal) solar irradiance series over the period from 1 Jul 2010 to 13 Jul 2010.

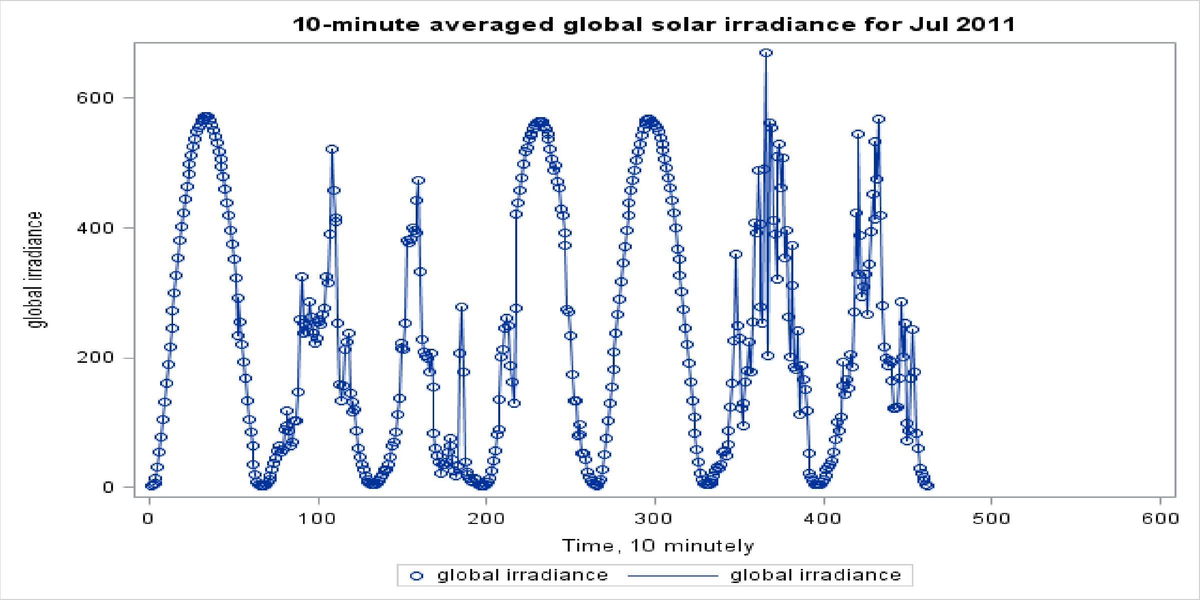

Figure 6.10: The plot of the 10-minutely averaged daylight global (horizontal) solar irradiance series over the period from 3 Jul 2011 to 9 Jul 2011.

6.4 SARIMA Models: Estimation and Forecasting

In this section, we fit the SARIMA models using the Box-Jenkins methodology, present SARIMA models fitted to each of the irradiance series as well as their prediction plots using SAS. It should be noted that a dotted or solid line on the graph indicates the start of multi-step forecasting. The in-sample diagnostics (e.g. AIC, BIC, R-squared and parsimony) and the hold-out sample prediction errors (e.g. MBE, MPE, MAPE and RMSE) are also provided for each of the models. Model estimation and residual analysis results are given in Appendix A and Appendix B respectively.

Clearly, all data series plots in Figures 6.3 to 6.10 exhibit seasonal variations. Therefore, Seasonal Autoregressive Integrated Moving Average (SARIMA) Models were deemed applicable for data series of this nature. We denote the irradiance variable by Y_t throughout. A seasonal differencing operator $(1 - L^S)^D$ was applied where necessary to transform the original series to a deseasonalized series $\Delta_S^D Y_t$. In the presence of the time-varying variability, the response variable was transformed by the Logarithmic transformation method. The best models were searched for by programming various candidate specifications in PROC ARIMA of the SAS Software, until the best were reached on the basis of methodologies discussed in Chapter 3.

The SARIMA model fitted to the 60-minutely Feb 2010 global horizontal irradiance series is given by

$$(1 + \phi_1 L)(1 + \Phi_{16} L^{16})(1 - L^{16}) \log(Y_t) = (1 + \theta_1 L + \Theta_{32} L^{32})\varepsilon_t. \qquad (6.3)$$

Parameter estimates of the SARIMA model given in Equation (6.3) can be found in Appendix A, Table A.1. All models were obtained via maximum likelihood (ML) as ML gives asymptotically normal estimates. The p-values in column 5 of Table A.1 are all less than a preset significance level, 0.05, which is an indication that all the model parameters are significant at a 5% level of significance. Results were obtained for all other fitted models and similar conclusions were made on the basis of the p-values for the parameters.

The residual analysis (Ljung-Box test) results of model in Equation (6.3) can be found in Appendix B, Table B.1. The output shows the p-values, in column 4, which are all greater than a preset significance level, 0.05. This indicates that the autocorrelation values are insignificant at a 5% level of significance and hence the residuals are uncorrelated up to higher lags (e.g. up to 48) and the white noise assumption is satisfied. The residual analysis was done for all other fitted SARIMA models in this section and similar conclusions can be made based on the p-values. Normality plots were done for residuals for all fitted SARIMA models and the normality assumption was not violated. The forecast plot for SARIMA model in Equation (6.3) is given in Figure 6.11 below.

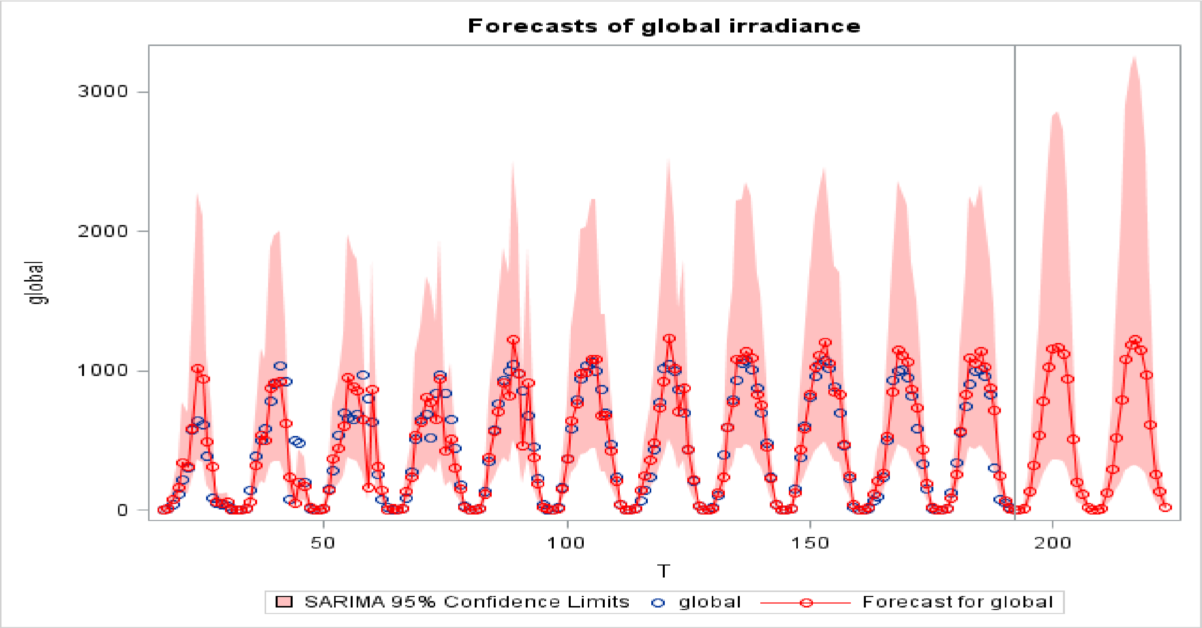

Figure 6.11: The plot of the actual versus predicted values for the 60-minutely averaged daylight global (horizontal) solar irradiance series from 2 Feb 2010 to 12 Feb 2010, plus two days ahead forecasting by model in Equation (6.3).

The SARIMA model fitted to the 60-minutely Feb 2011 global horizontal irradiance series is given by

$$(1+\phi_3 L^3+\phi_{12}L^{12}+\Phi_{14}L^{14})(1-L^{14})\log(Y_t)=(1+\theta_1 L+\theta_2 L^2+\Theta_{28}L^{28})\varepsilon_t. \quad (6.4)$$

Parameter estimates of the SARIMA model given in Equation (6.4) can be found in Appendix A, Table A.2, and the residual analysis (Ljung-Box test) results in Appendix B, Table B.2. The forecast plot for this model is given in Figure 6.12 below.

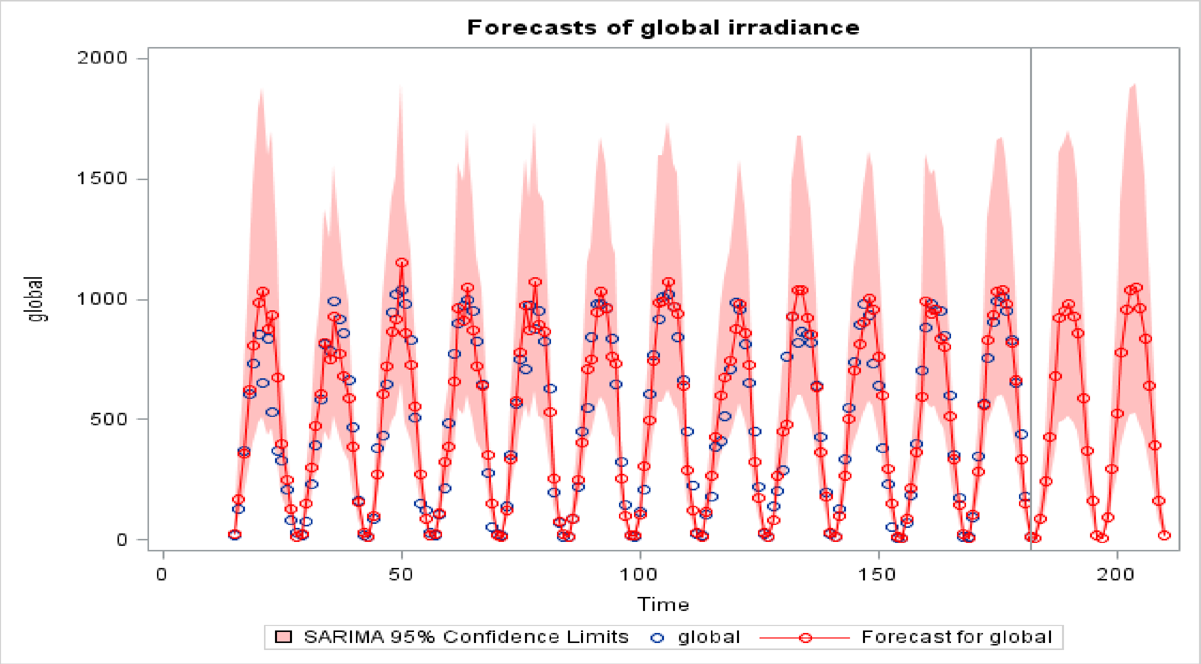

Figure 6.12: The plot of the actual versus predicted values for the 60-minutely averaged daylight global (horizontal) solar irradiance series from 2 Feb 2011 to 13 Feb 2011, plus two days ahead forecasting by model in Equation (6.4).

The SARIMA model fitted to the 60-minutely Jul 2010 global horizontal irradiance series is given by

$$(1 + \phi_1 L + \phi_2 L^2)(1 - L^{12})\log(Y_t) = (1 + \Theta_{12} L^{12})\varepsilon_t. \qquad (6.5)$$

Parameter estimates of the SARIMA model given in Equation (6.5) can be found in Appendix A, Table A.3, and the residual analysis (Ljung-Box test) results in Appendix B, Table B.3. The forecast plot for this model is given in Figure 6.13 below.

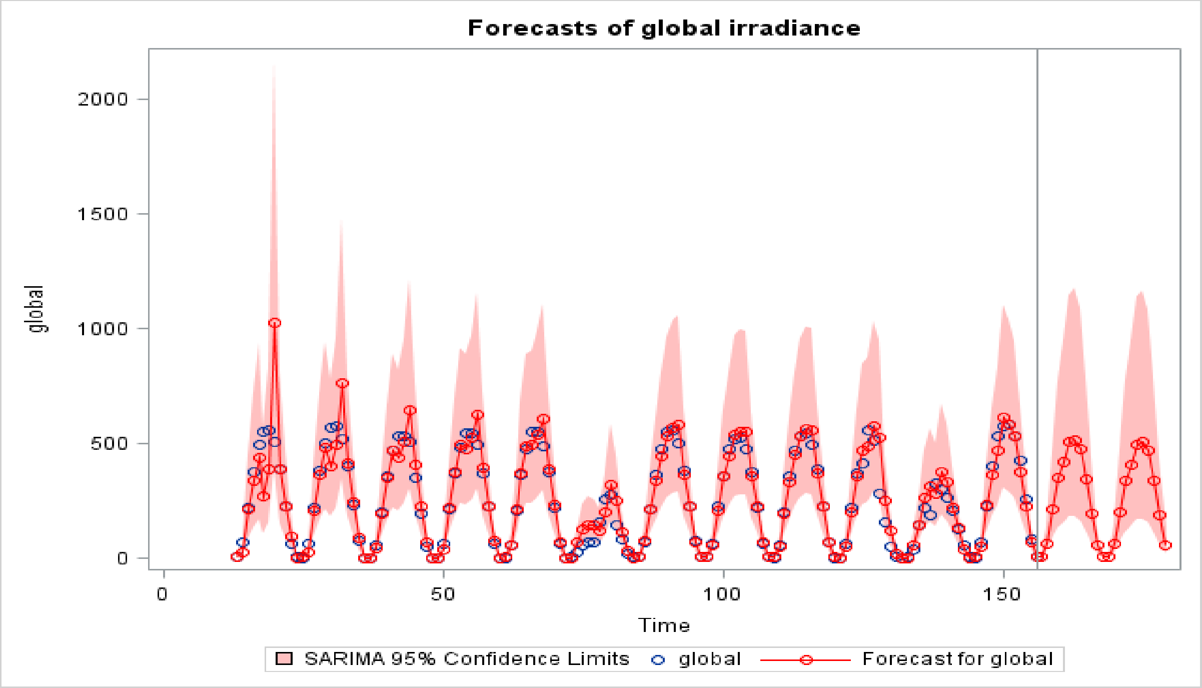

Figure 6.13: The plot of the actual versus predicted values for the 60-minutely averaged daylight global (horizontal) solar irradiance series from 2 July 2010 to 13 Jul 2010, plus two days ahead forecasting by model in Equation (6.5).

The SARIMA model D, fitted to the 60-minutely Jul 2011 global horizontal irradiance series is given by

$$(1+\phi_1 L+\phi_2 L^2+\phi_{12} L^{12}+\phi_{15} L^{15}+\Phi_{24} L^{24}+\Phi_{36} L^{36})(1-L^{36})Y_t=(1+\phi_1 L+\phi_2 L^2)\varepsilon_t. \tag{6.6}$$

Parameter estimates of the SARIMA model given in Equation (6.6) can be found in Appendix A, Table A.4, and the residual analysis (Ljung-Box test) results in Appendix B, Table B.4. The forecast plot for this model is given in Figure 6.14 below.

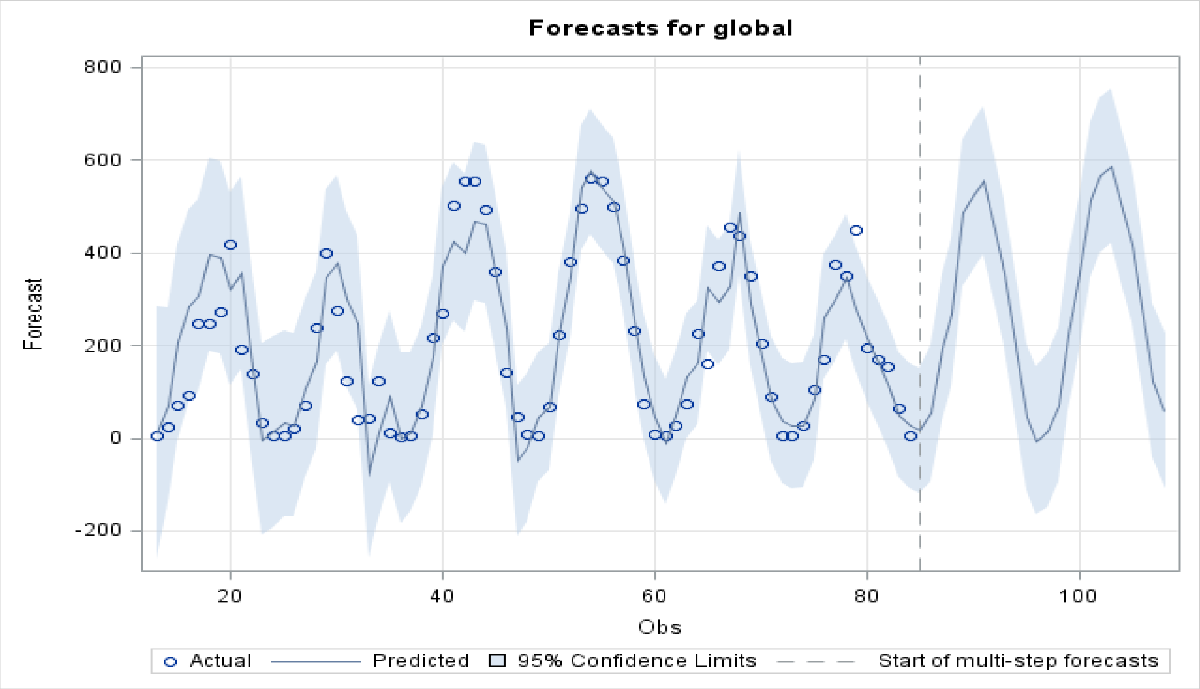

Figure 6.14: The plot of the actual versus predicted values for the 60-minutely averaged daylight global (horizontal) solar irradiance series from 2 July 2011 to 7 July 2011, plus two days ahead forecasting by model in Equation (6.6).

The SARIMA model fitted to the 10-minutely Feb 2010 global horizontal irradiance series is given by

$$(1+\phi_1 L+\phi_{17}L^{17})(1-L^{90})\log(Y_t) = (1+\theta_1 L+\theta_3 L^3+\theta_{10}L^{10}+\theta_{13}L^{13})(1+\Theta_{90}L^{90})\varepsilon_t. \tag{6.7}$$

Parameter estimates of the SARIMA model given in Equation (6.7) can be found in Appendix A, Table A.5, and the residual analysis (Ljung-Box test) results in Appendix B, Table B.5. The forecast plot for this model is given in Figure 6.15 below.

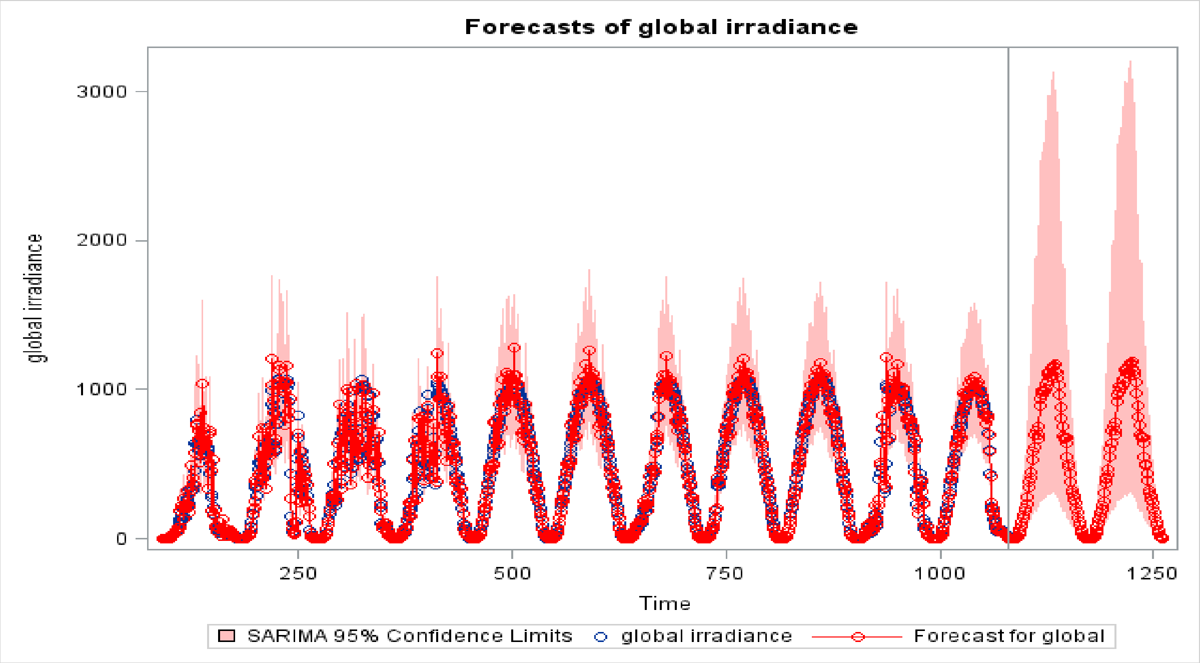

Figure 6.15: The plot of the actual versus predicted values for the 10-minutely averaged daylight global (horizontal) solar irradiance series from 2 Feb 2010 to 12 Feb 2010, plus two days ahead forecasting by model in Equation (6.7).

The SARIMA model fitted to the 10-minutely Feb 2011 global horizontal irradiance series is given by

$$(1+\phi_1 L+\phi_2 L^2+\phi_3 L^3+\phi_6 L^6+\phi_{10} L^{10}+\phi_{11} L^{11})(\Theta_{84} L^{84}+\Theta_{168} L^{168})(1-L^{84})\log(Y)=$$

$$(1+\theta_2 L^2+\theta_4 L^4)\varepsilon_t. \tag{6.8}$$

Parameter estimates of the SARIMA model given in Equation (6.8) can be found in Appendix A, Table A.6, and the residual analysis (Ljung-Box test) results in Appendix B, Table B.6. The forecast plot for this model is given in Figure 6.16 below.

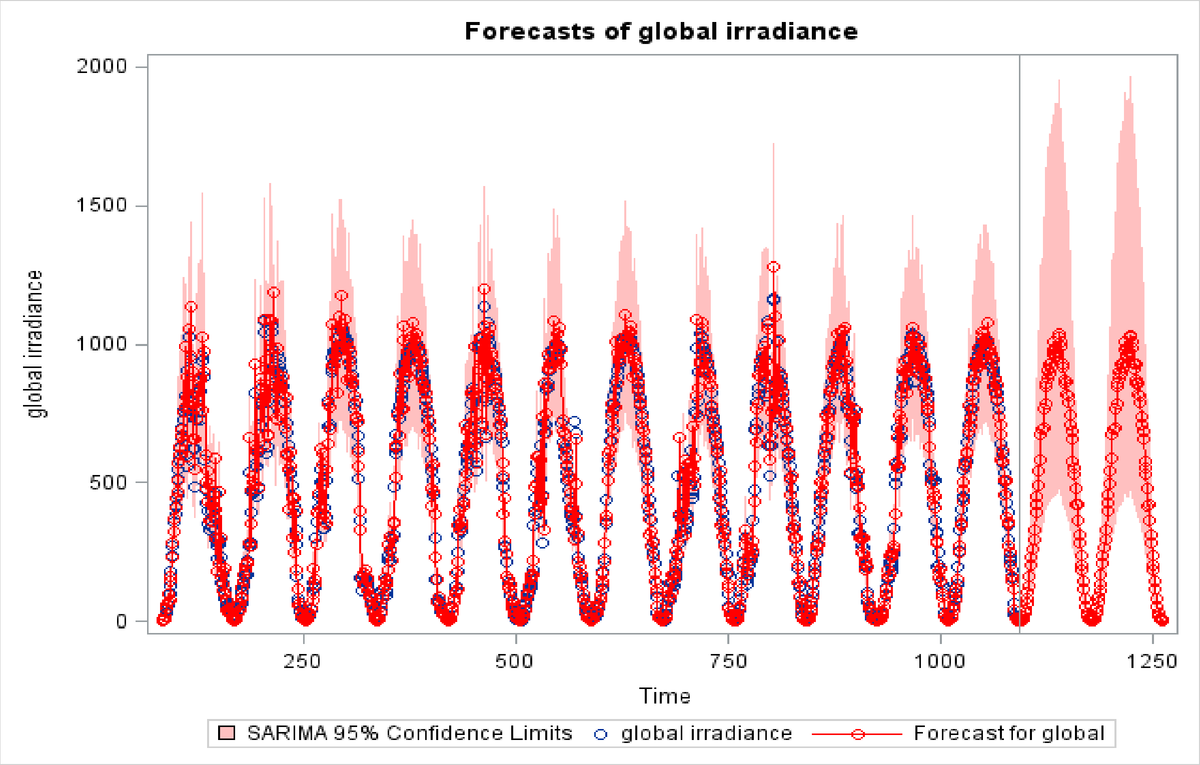

Figure 6.16: The plot of the actual versus predicted values for the 10-minutely averaged daylight global (horizontal) solar irradiance series from 2 Feb 2011 to 13 Feb 2011, plus two days ahead forecasting by model in Equation (6.8).

The SARIMA model fitted to the 10-minutely Jul 2010 global horizontal irradiance series is given by

$$(1+\theta_1 L+\theta_2 L^2)(1+\Theta_{68}L^{68})(1-L^{68})\log(Y_t) = (1+\theta_2 L^2)\varepsilon_{t.} \tag{6.9}$$

Parameter estimates of the SARIMA model given in Equation (6.9) can be found in Appendix A, Table A.7, and the residual analysis (Ljung-Box test) results in Appendix B, Table B.7. The forecast plot for this model is given in Figure 6.17 below.

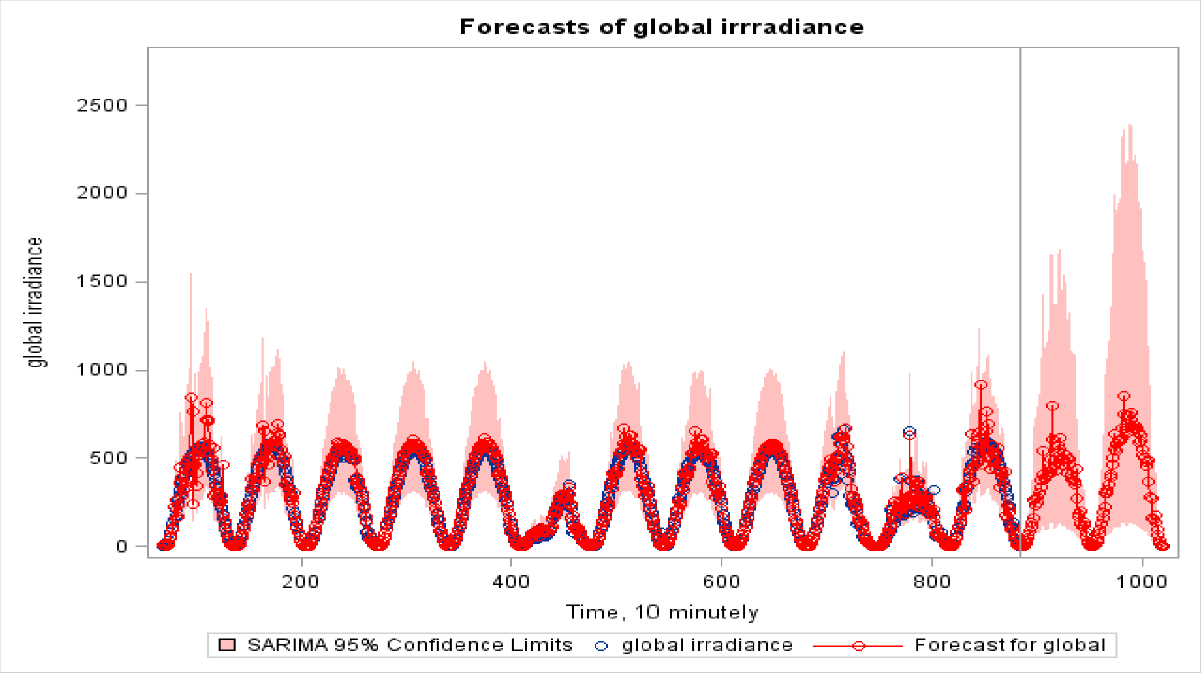

Figure 6.17: The plot of the actual versus predicted values for the 10-minutely averaged daylight global (horizontal) solar irradiance series from 1 Jul 2010 to 13 Jul 2010, plus two days ahead forecasting by model in Equation (6.9).

The SARIMA model fitted to the 10-minutely Jul 2011 global horizontal irradiance series is given by

$$(1-\phi_1 L+\phi_7 L^7-\phi_{11}L^{11}-\phi_{12}L^{12}-\phi_{14}L^{14})(1-\Phi_{66}L^{66}-\Phi_{132}L^{132})(1-L^{66})Y_t=\varepsilon_t. \tag{6.10}$$

Parameter estimates of the SARIMA model given in Equation (6.10) can be found in Appendix A, Table A.8, and the residual analysis (Ljung-Box test) results in Appendix B, Table B.8. The forecast plot for this model is given in Figure 6.18 below.

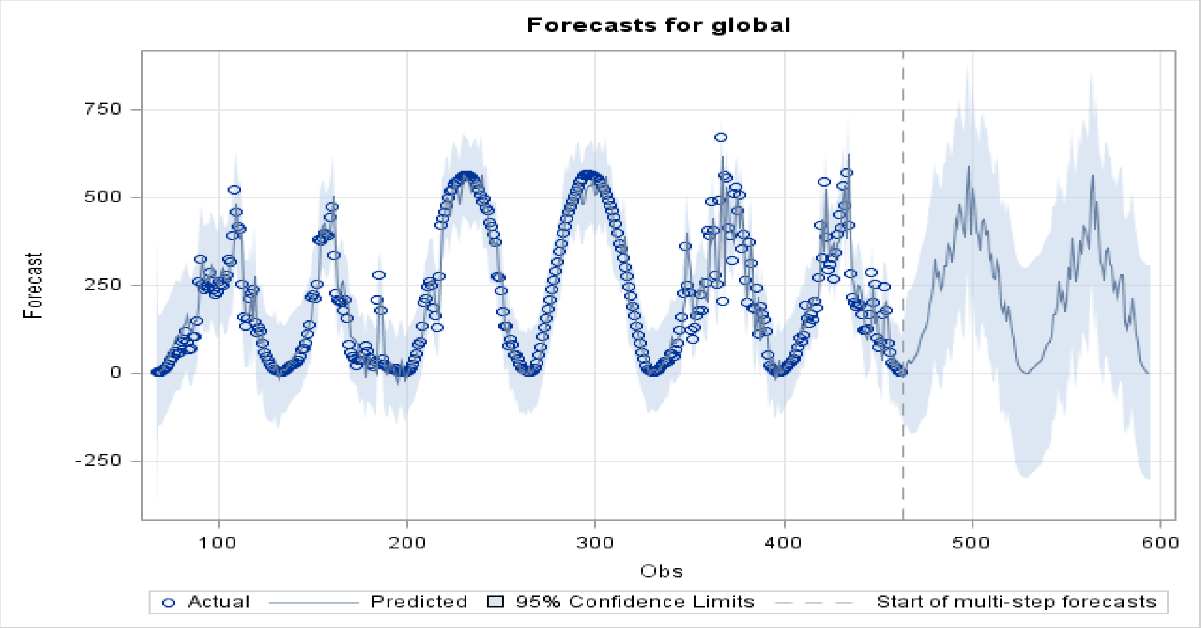

Figure 6.18: Plot of the actual versus predicted values for the 10-minutely averaged daylight global solar irradiance series from 4 Jul 2011 to 11 Jul 2011, plus two days ahead forecasting by model in Equation (6.10).

In Tables 6.5 to 6.8, are the in-sample diagnostics and out-of-sample prediction errors for each of the SARIMA models fitted. These statistical values indicate the goodness of fit in terms of the parsimony (AIC and BIC), determined by the number of parameters in the model, as well as the coefficient of determination. It is commonly known that the smaller the magnitudes of each of these indicators the better the fit. Table 6.6 and 6.8 clearly show that the SARIMA models fitted to both 10-minutely and 60-minutetly data provide better forecasts in summer than winter for the year 2010. It is hard to conclude the same for the year 2011 as the pattern becomes unclear.

Table 6.5: AIC, SBC, R-squared and parsimony for the SARIMA models fitted to the 60-minutely averaged global (horizontal) irradiance series.

			In-sample Model Section Diagnostics			
Model	**Season**		**AIC**	**SBC**	**R-squared**	**Parsimony**
Eq. (6.3)	**Summer**	Feb 2010	195.738	208.420	0.959	4
Eq. (6.4)		Feb 2011	1973.942	1992.686	0.946	6
Eq. (6.5)	**Winter**	Jul 2010	1490.317	1499.226	0.957	3
Eq. (6.6)		Jul 2011	841.435	857.372	0.840	7

Table 6.6: Forecast accuracy measures for the SARIMA models fitted on each of the 60-minutely averaged daylight global (horizontal) solar irradiance series.

			Model forecast accuracy measure			
Model	**Season**		**MBE**	**MPE**	**MAPE**	**RMSE**
Eq. (6.3)	**Summer**	Feb 2010	185.460	101.753	109.055	286.207
Eq. (6.4)		Feb 2011	39.757	10.866	50.641	143.673
Eq. (6.5)	**Winter**	Jul 2010	-55.230	-37.414	37.414	64.722
Eq. (6.6)		Jul 2011	-31.802	-26.321	63.111	45.935

Table 6.7: In-sample diagnostics for the SARIMA models fitted on each of the 10-minutely averaged daylight global (horizontal) solar irradiance series**.**

			In-sample Model Section Diagnostics			
Model	**Season**		**AIC**	**SBC**	**R-square**	**Parsimony**
Eq. (6.7)	**Summer**	Feb 2010	-162.283	-123.101	0.986	8
Eq. (6.8)		Feb 2011	11613.350	11662.500	0.958	10
Eq. (6.9)	**Winter**	Jul 2010	400.857	419.660	0.966	4
Eq. (6.10)		Jul 2011	4381.669	4409.486	0.911	7

Table 6.8: Prediction errors for the SARIMA models fitted on each of the 10-minutely averaged global (horizontal) irradiance series.

			Model forecast accuracy measure			
Model	**Season**		**MBE**	**MPE**	**MAPE**	**RMSE**
Eq. (6.7)	**Summer**	Feb 2010	165.947	130.779	138.932	267.011
Eq. (6.8)		Feb 2011	38.075	25.119	44.692	155.747
Eq. (6.9)	**Winter**	Jul 2010	19.743	18.694	30.123	84.540
Eq. (6.10)		Jul 2011	-94.178	-28.119	38.167	122.249

6.5 Spectral Analysis

In this section we present the spectral analysis results for the test for the existence of periodicities in the data using frequency domain techniques discussed in Chapter 3. We also make use of the F-test with the test statistic given in Equation (3.64) to test for the statistical significance of the largest periodogram ordinates at the 5% level of significance (see Table 6.9). In Figures 6.3 to 6.8, the periodogram plots for all of the irradiance series are presented. The analysis results show that there are periodicities in all of eight data series under investigation. The single strongest spikes corresponding to largest periods testifies to the apparent day cycles (seasonalities) in all data series under study. The *Bartlett's Kolmogorov-Smirnov Statistic* for each series, uniform (0,1), shows that generally the series is not white noise. In the next section we present the results of HCSARIMA models for eight data series.

Table 6.9: Periodogram analysis for all data sets and F-test for the significance of the largest ordinates.

Series	Obs	λ_k	Period k	Max $I(\lambda_k)$	F-statistic	F-critical
60-min Feb 2011	14	0.449	14.00	263.305	258.825	3.046
60-min Jul 2011	8	0.524	12.000	2027789.760	98.822	3.109
10-min Feb 2011	14	0.075	84.000	126158728.780	4656.780	3.004
10-min Jul 2011	8	0.095	66.000	10444739.500	403.575	3.015
60-min Feb 2010	13	0.393	16.000	585.612	311.408	3.044
60-min Jul 2010	14	0.524	12.000	324.676	183.773	3.055
10-min Feb 2010	13	0.070	90.000	2675.835	1265.337	3.004
10-min Jul 2010	14	0.092	68.000	1493.828	785.041	3.006

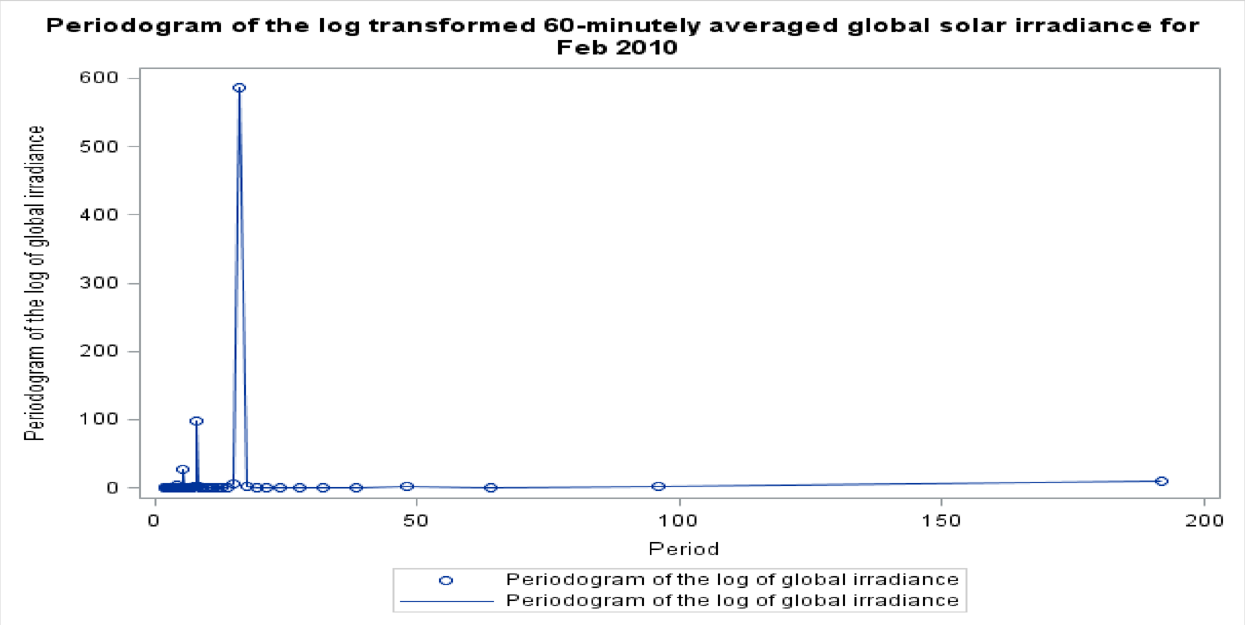

Figure 6.19: Periodogram plot of the log transformed 60-minutely averaged irradiance series for the period of the 1st to the 12th for Feb 2010.

Figure 6.19 shows the largest ordinate at period 16, corresponding to the harmonic frequency $2\pi/16$. The Fisher's Kappa statistic is equal to 72.883, which is significant at the 1% level of significance indicating that the largest ordinate is highly significant.

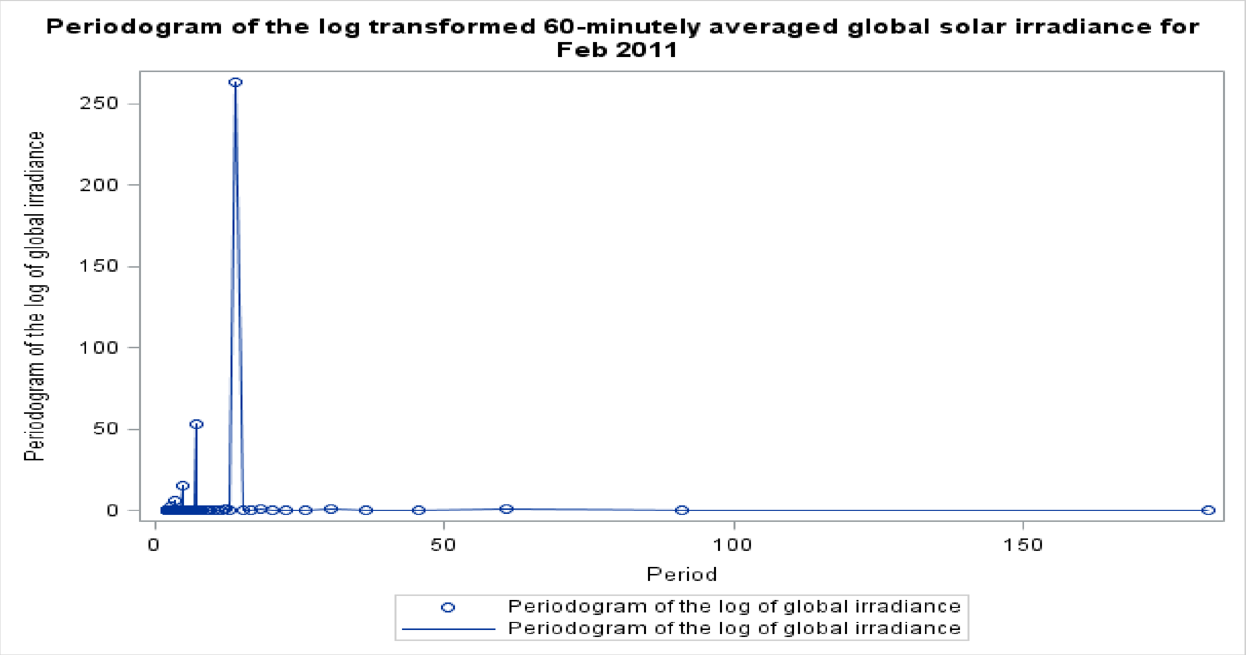

Figure 6.20: Periodogram plot for the log transformed 60-minutely averaged irradiance series for the period of the 1st to the 13th for Feb 2011.

Figure 6.20 shows the largest ordinate at period 14, corresponding to the harmonic frequency $2\pi/14$. The Fisher's Kappa statistic is equal to 66.875, which is significant at the 1% level of significance indicating that the largest ordinate is highly significant.

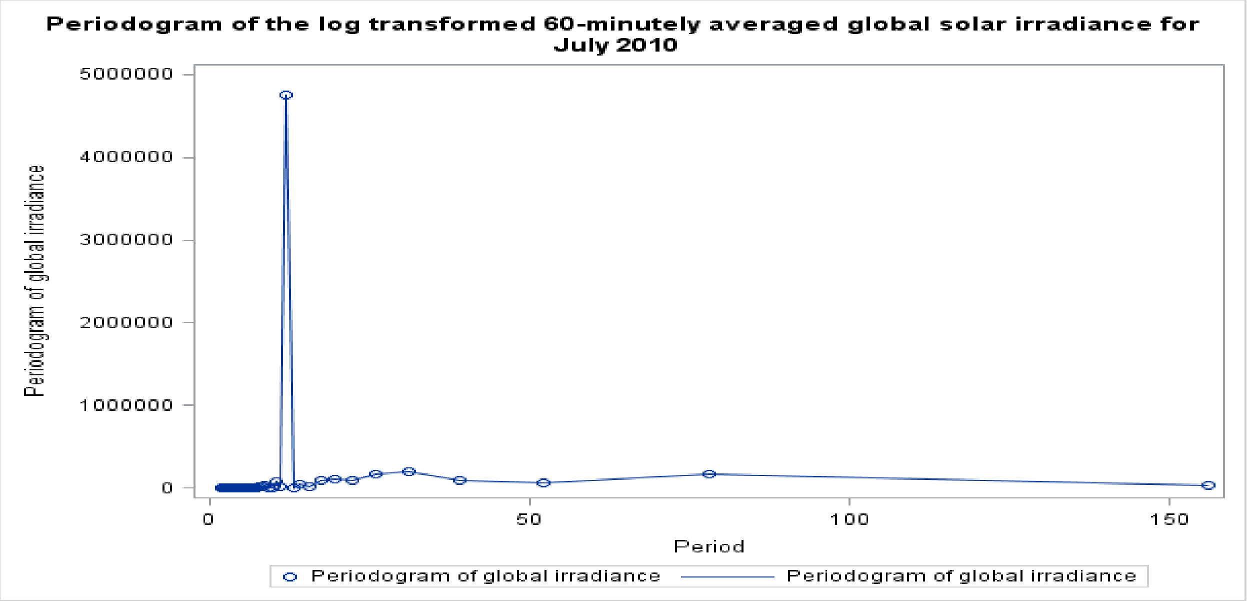

Figure 6.21: Periodogram plot for the log transformed 60-minutely averaged irradiance series for the period of the 1st to the 13th for Jul 2010.

Figure 6.21 shows the largest ordinate at period 12, corresponding to the harmonic frequency $2\pi/12$. The Fisher's Kappa statistic is equal to 54.368, which is significant at the 1% level of significance indicating that the largest ordinate is highly significant.

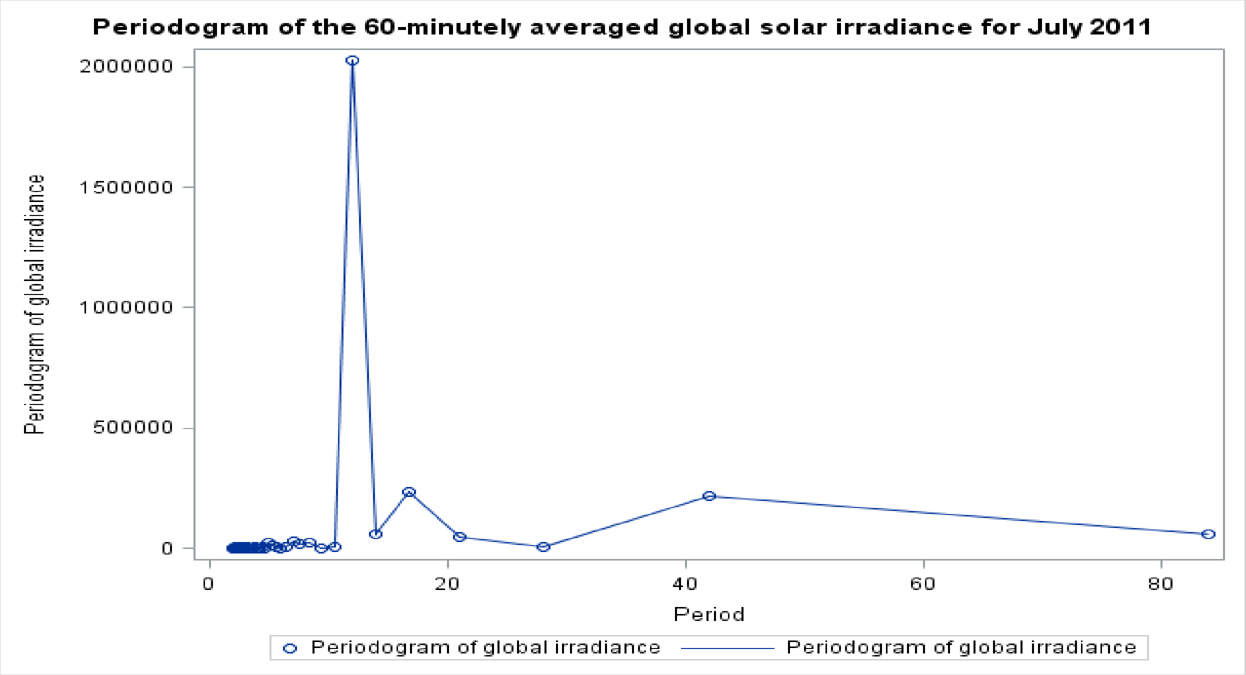

Figure 6.22: Periodogram plot for the 60-minutely averaged irradiance series for the period of the 3rd to the 9th for Jul 2011.

Figure 6.22 shows the largest ordinate at period 12, corresponding to the harmonic $2\pi/12$. The Fisher's Kappa statistic is equal to 29.082, which is significant at the 1% level of significance indicating that the largest ordinate is highly significant.

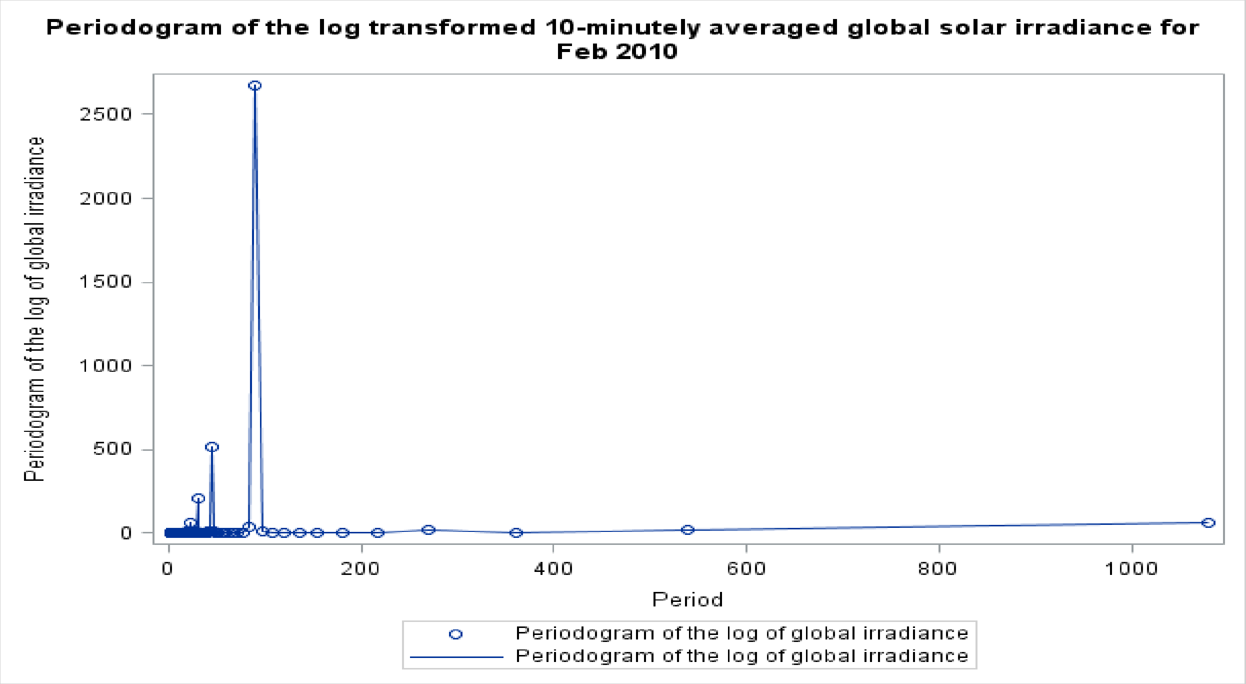

Figure 6.23: Periodogram plot for the log transformed10-minutely averaged irradiance series for the period of the 1st to the 12th for Feb 2010.

Figure 6.23 shows the largest ordinate at period 90, corresponding to the harmonic frequency $2\pi/90$. The Fisher's Kappa statistic is equal to 378.092, which is significant at the 1% level of significance indicating that the largest ordinate is highly significant.

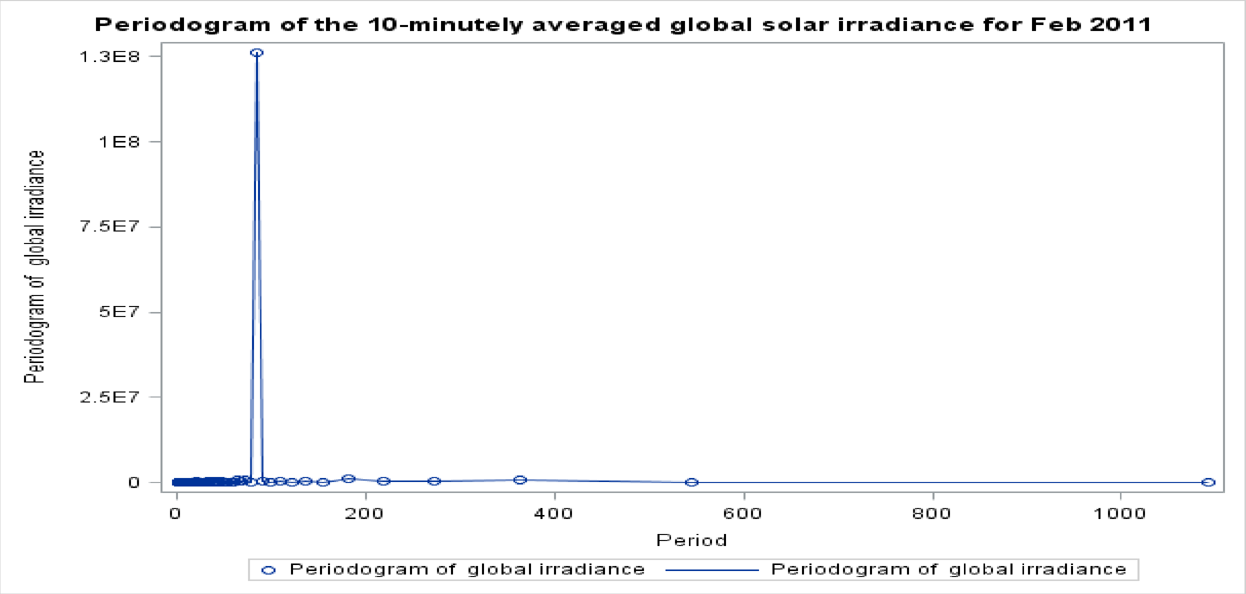

Figure 6.24: Periodogram plot of the 10-minutely averaged irradiance series for the period of the 1st to the 13th for Feb 2011.

Figure 6.24 shows the largest ordinate at period 84, corresponding to the harmonic frequency $2\pi/84$. The Fisher's Kappa statistic is equal to 487.970, which is significant at the 1% level of significance indicating that the largest ordinate is highly significant.

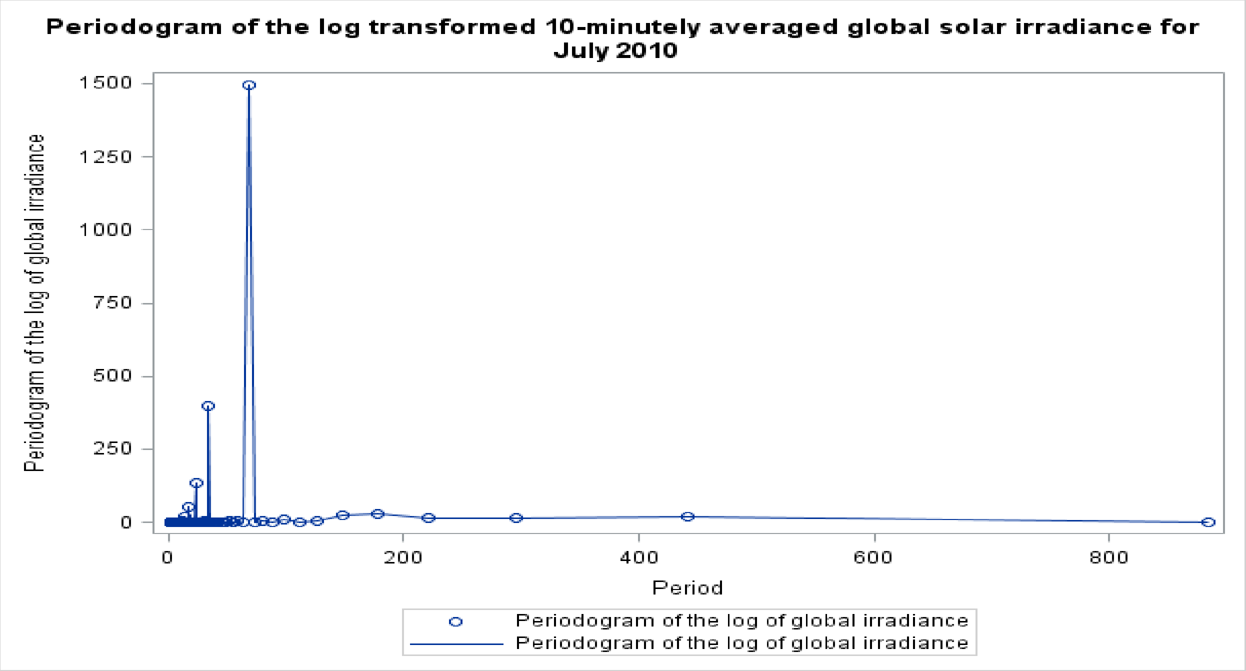

Figure 6.25: Periodogram plot for the log transformed 10-minutely averaged irradiance series for the period of the 1st to the 13th for Jul 2010.

Figure 6.25 shows the largest ordinate at period 68, corresponding to the harmonic frequency $2\pi/68$.The Fisher's Kappa statistic is equal to 282.490, which is significant at the 1% level of significance indicating that the largest ordinate is highly significant.

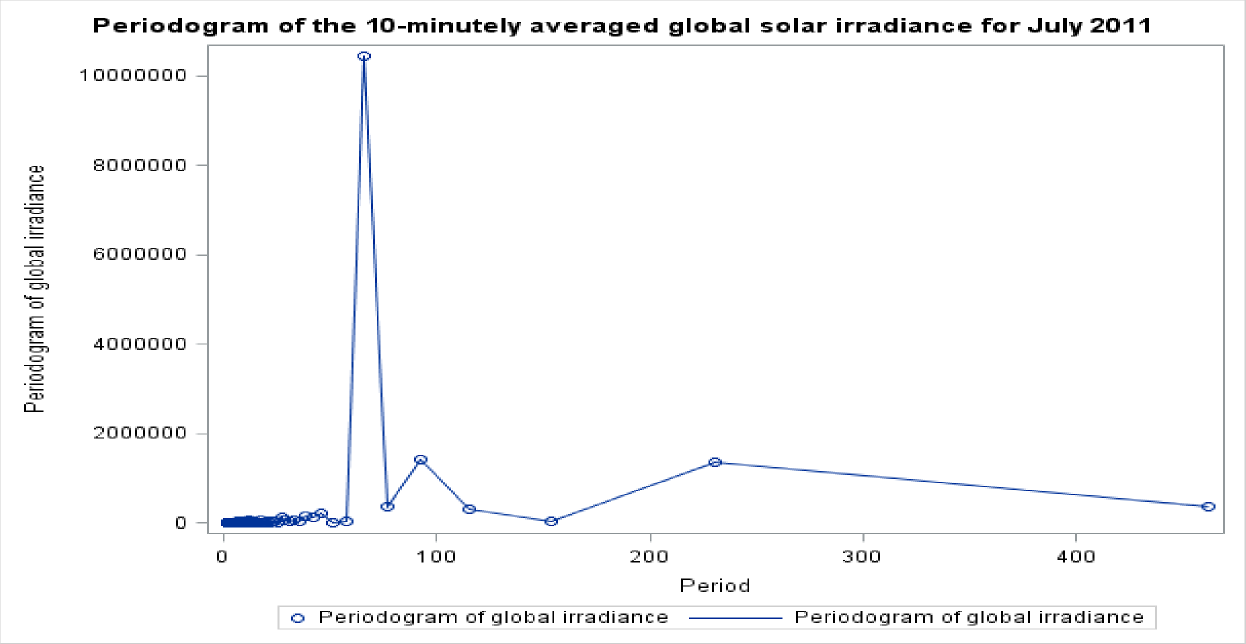

Figure 6.26: Periodogram plot of the 10-minutely averaged irradiance series for the period of the 3rd to the 9th for Jul 2011.

Figure 6.26 shows the largest ordinate at period 66, corresponding to the harmonic frequency $2\pi/66$. The Fisher's Kappa statistic is equal to 146.621, which is significant at the 1% level of significance indicating that the largest ordinate is highly significant.

6.6 HCSARIMA Models: Estimation and Residual Analysis

In this section, we present all the HCSARIMA models fitted to the irradiance series, given in Figures 6.3 to 6.10, using SAS program, as well as their prediction and forecast plots. The in-sample diagnostics (e.g. AIC, BIC, R-squared and parsimony) and the hold-out sample prediction errors (e.g. MBE, MPE, MAPE and RMSE) are also provided for each of these models. Model estimation and residual analysis results are also presented in Appendix A and Appendix B respectively. As in SARIMA modelling, similar analysis was carried out for the HCSARIMA class of models. Using the same methods (i.e. ML for parameter estimation and Ljung-Box test for residuals), the same conclusions, based on the SAS outputs, were made for all the models of this class given in this section.

The HCSARIMA model fitted to the 60-minutely Feb 2010 global horizontal irradiance series is given by

$$Y_t = \mu + bt + \alpha \cos\left(\frac{2\pi}{16}\right) t + \beta \sin\left(\frac{2\pi}{16}\right) + (1 + \phi_1 L + \phi_2 L^2 + \Phi_{16} L^{16})\varepsilon_t + e_t. \qquad (6.11)$$

Parameter estimates of the HCSARIMA model given in Equation (6.11) can be found in Appendix A, Table A.9, and the residual analysis (Ljung-Box test) results in Appendix B, Table B.9. The forecast plot for this model is given in Figure 6.27 below.

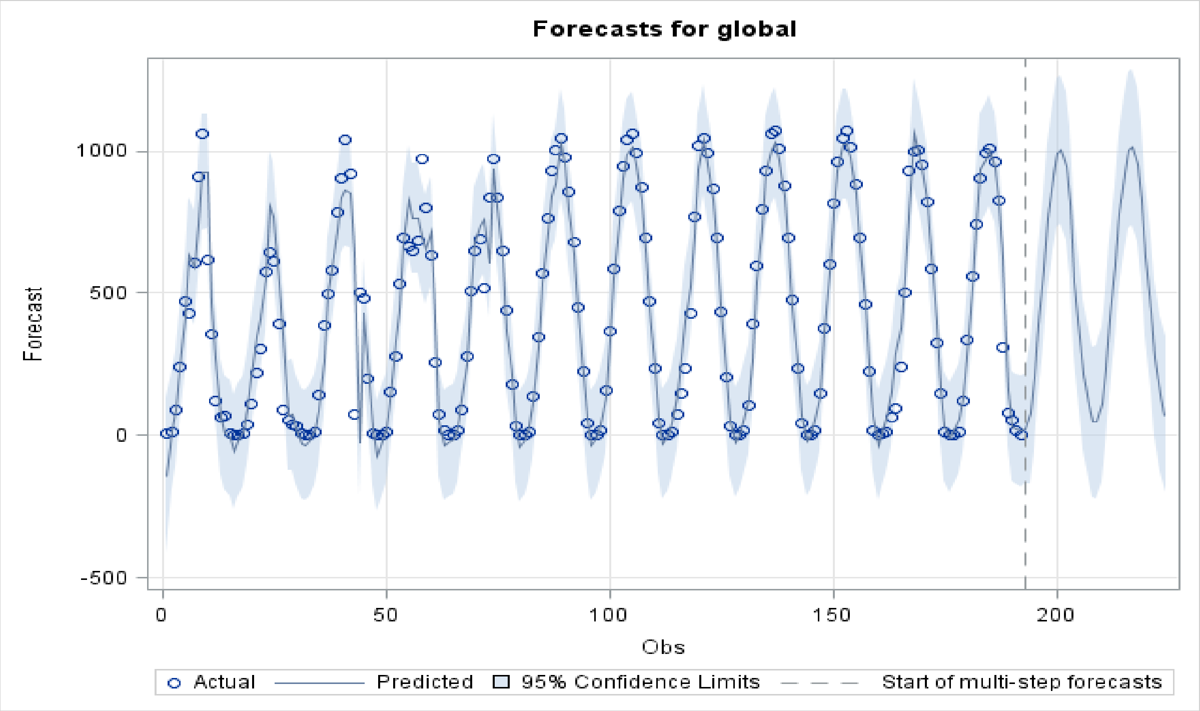

Figure 6.27: Plot of the actual versus predicted values for the 60-minutely averaged daylight global (horizontal) solar irradiance series from 1 Feb 2010 to 12 Feb 2010, plus two days ahead forecasting by model in Equation (6.11).

The HCSARIMA model fitted to the 60-minutely Feb 2011 global horizontal irradiance series is given by

$$Y_t = \mu + \alpha_1 \cos\left(\frac{2\pi}{14}\right) t + \beta_1 \sin\left(\frac{2\pi}{14}\right) + \alpha_2 \cos\left(\frac{4\pi}{14}\right) + (1 + \phi_1 L + \Phi_{56} L^{56})\varepsilon_t + e_t. \quad (6.12)$$

Parameter estimates of the HCSARIMA model given in Equation (6.12) can be found in Appendix A, Table A.10, and the residual analysis (Ljung-Box test) results in Appendix B, Table B.10. The forecast plot for this model is given in Figure 6.28 below.

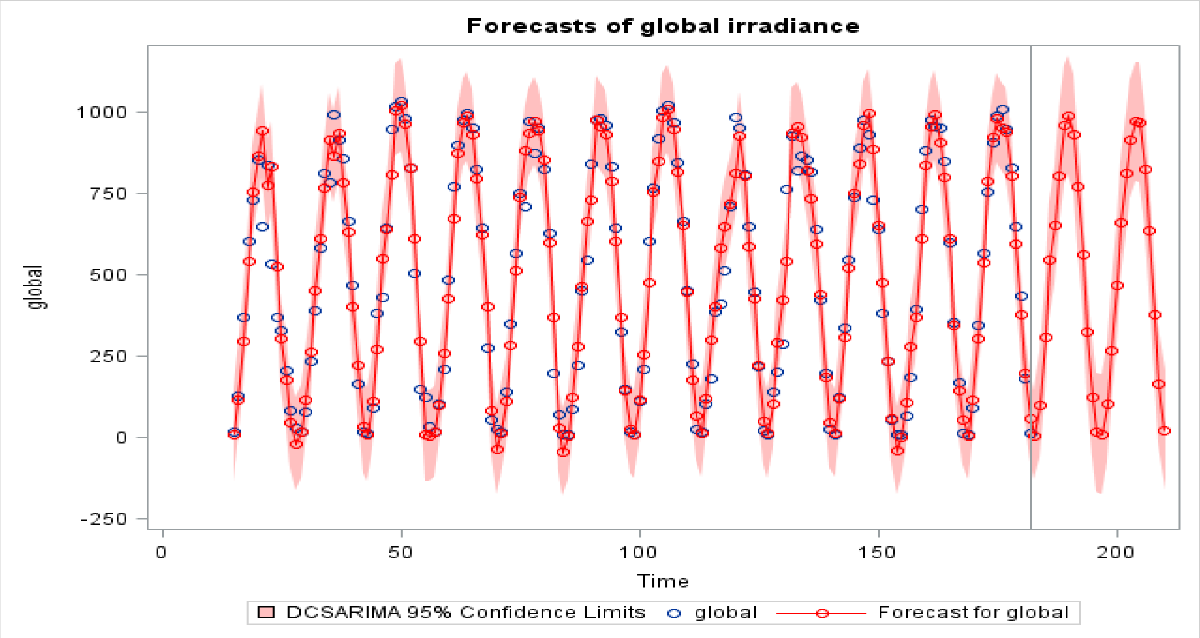

Figure 6.28: Plot of the actual versus predicted values for the 60-minutely averaged daylight global (horizontal) solar irradiance series from 2 Feb 2011 to 13 Feb 2011, plus two days ahead forecasting by model given in Equation (6.12).

The HCSARIMA model fitted to the 60-minutely Jul 2010 global horizontal irradiance series is given by

$$Y_t = \mu + \alpha \cos\left(\frac{2\pi}{12}\right) t + \beta \sin\left(\frac{2\pi}{12}\right) + (1 + \phi_1 L)(1 + \theta_1 L)\varepsilon_t + e_{t.} \qquad (6.13)$$

Parameter estimates of the HCSARIMA model given in Equation (6.13) can be found in Appendix A, Table A.10, and the residual analysis (Ljung-Box test) results in Appendix B, Table B.10. The forecast plot for this model is given in Figure 6.29 below.

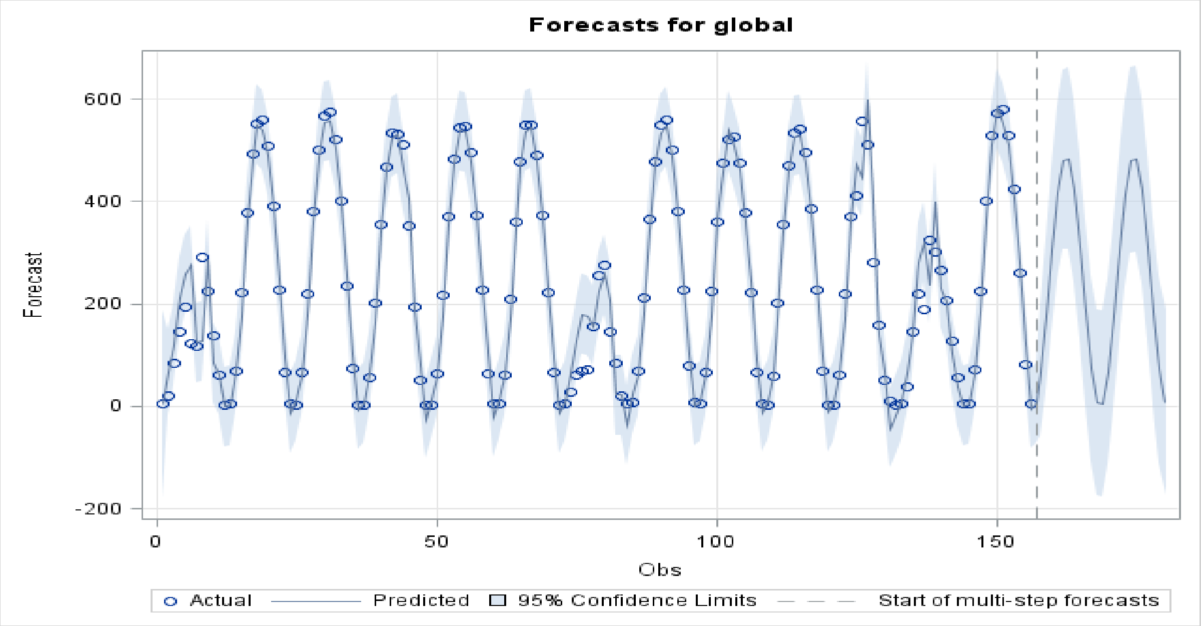

Figure 6.29: Plot of the actual versus predicted values for the 60-minutely averaged global (horizontal) solar irradiance series from 1 Jul 2010 to 13 Jul 2010, plus two days ahead forecasting by model in Equation (6.13).

The HCSARIMA model fitted to the 60-minutely Jul 2011 global horizontal irradiance series is given by

$$Y_t = \mu + \alpha \cos\left(\frac{2\pi}{12}\right) t + \beta \sin\left(\frac{2\pi}{12}\right) + (1 + \phi_1 L + \Phi_{15} L^{15})\varepsilon_t + e_t \,. \qquad (6.14)$$

Parameter estimates of the HCSARIMA model given in Equation (6.14) can be found in Appendix A, Table A.11, and the residual analysis (Ljung-Box test) results in Appendix B, Table B.11. The forecast plot for this model is given in Figure 6.30 below.

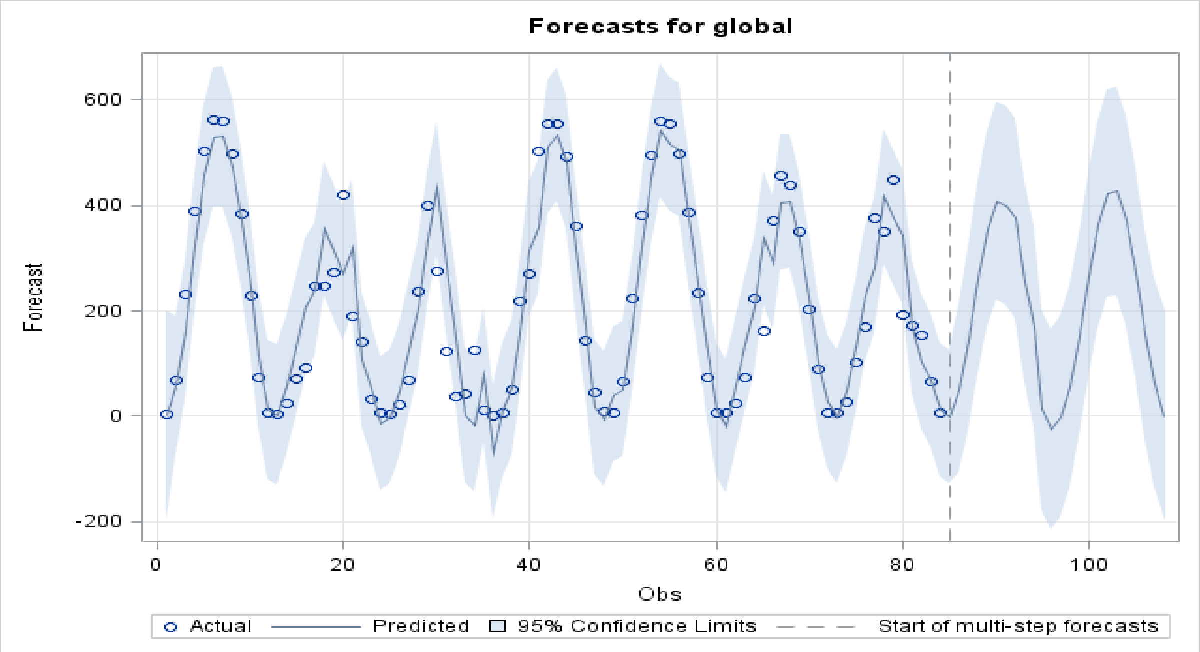

Figure 6.30: Plot of the actual versus predicted values for the 60-minutely averaged daylight global (horizontal) solar irradiance series from 3 Jul 2011 to 9 Jul 2011, plus two days ahead forecasting by model in Equation (6.14).

The HCSARIMA model E1, fitted to the 10-minutely Feb 2010 irradiance series is given by

$$\log(Y_t) = \mu + \alpha_1 \cos\left(\frac{2\pi}{90}\right) t + \beta_1 \sin\left(\frac{2\pi}{90}\right) + \alpha_2 \cos\left(\frac{4\pi}{90}\right) t + \beta_2 \sin\left(\frac{4\pi}{90}\right) + (1 + \phi_1 L + \phi_2 L^2 + \phi_{11} L^{11} + \phi_{21} L^{21} + \Phi_{55} L^{55} + \Phi_{81} L^{81} + \Phi_{90} L^{90} + \Phi_{92} L^{92} + \Phi_{93} L^{93})(\theta_1 L + \theta_2 L^2 + \Theta_{88} L^{88} + \Theta_{89} L^{89})\varepsilon_t + e_t. \quad (6.15)$$

Parameter estimates of the HCSARIMA model given in Equation (6.15) can be found in Appendix A, Table A.13, and the residual analysis (Ljung-Box test) results in Appendix B, Table B.13. The forecast plot for this model is given in Figure 6.31 below.

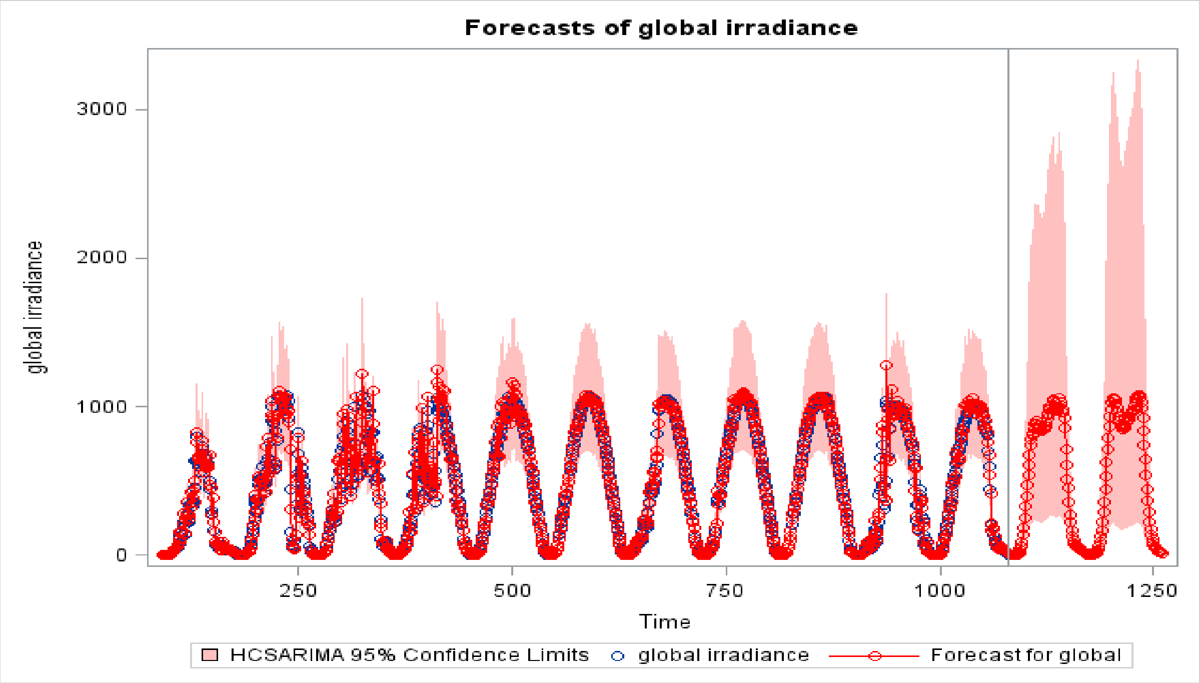

Figure 6.31: Plot of the actual versus predicted values for the 10-minutely averaged daylight global (horizontal) solar irradiance series from 2 Feb 2010 to 12 Feb 2010, plus two days ahead forecasting by model in Equation (6.15).

The HCSARIMA model fitted to the 10-minutely Feb 2011 global horizontal irradiance series is given by

$$Y_t = \mu + \alpha_1 \cos\left(\frac{2\pi}{84}\right) t + \beta_1 \sin\left(\frac{2\pi}{84}\right) + (1 + \phi_1 L + \phi_2 L^2 + \phi_3 L^3 + \phi_6 L^6 + \phi_7 L^7)\varepsilon_t + e_t. \quad (6.16)$$

Parameter estimates of the HCSARIMA model given by Equation (6.16) can be found in Appendix A, Table A.14, and the residual analysis (Ljung-Box test) results in Appendix B, Table B.14. The forecast plot for this model is given in Figure 6.32 below.

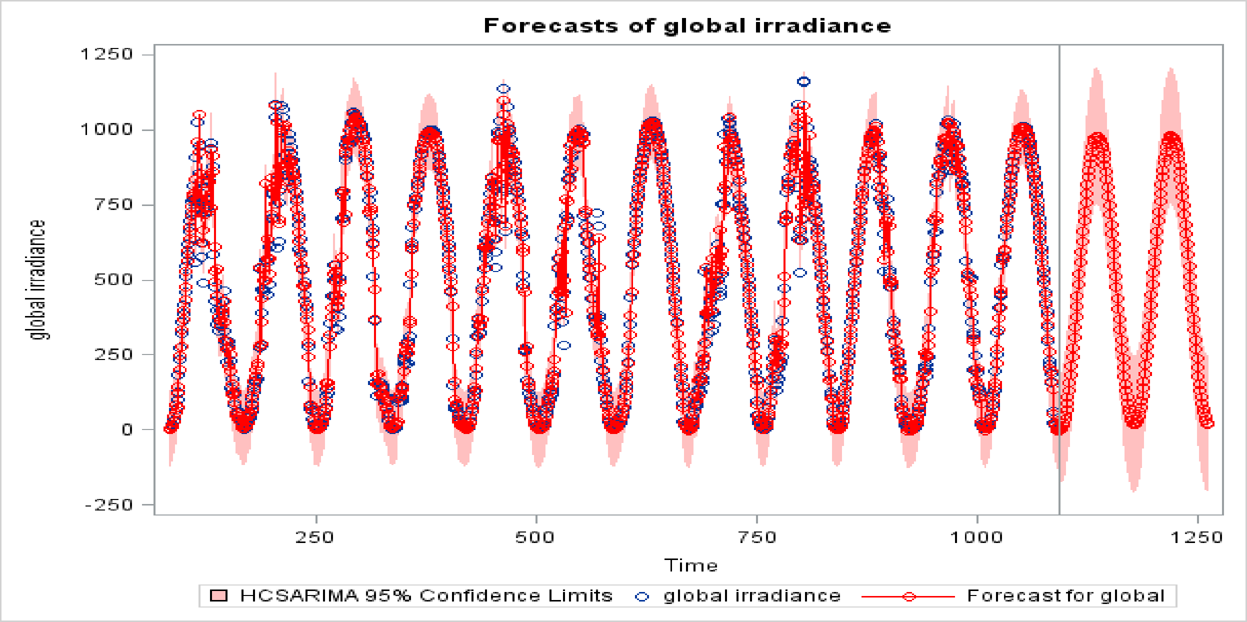

Figure 6.32: Plot of the actual versus predicted values for the 10-minutely averaged daylight global (horizontal) solar irradiance series from 2 Feb 2011 to 13 Feb 2011, plus two days ahead forecasting by model in Equation (6.16).

The HCSARIMA model fitted to the 10-minutely Jul 2010 global horizontal irradiance series is given by

$$\log(Y_t) = \mu + \alpha_1 \cos\left(\frac{2\pi}{68}\right) t + \alpha_2 \cos\left(\frac{4\pi}{68}\right) + (1 + \phi_1 L + \phi_3 L^3 + \phi_{17} L^{17} + \Phi_{68} L^{68})(1 + \theta_{68} L^{68})\varepsilon_t + e_t. \qquad (6.17)$$

Parameter estimates of the HCSARIMA model given in Equation (6.17) can be found in Appendix A, Table A.15, and the residual analysis (Ljung-Box test) results in Appendix B, Table B.15. The forecast plot for this model is given in Figure 6.33 below.

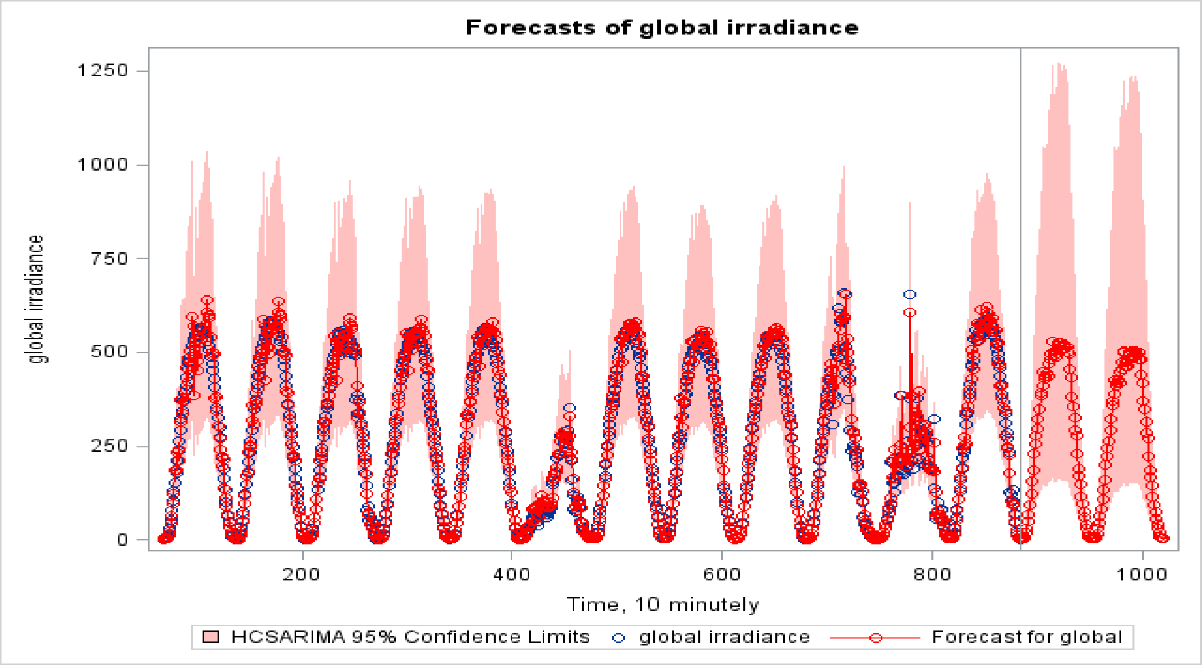

Figure 6.33: Plot of the actual versus predicted values for the 10-minutely averaged daylight global (horizontal) solar irradiance series from 2 Jul 2010 to 13 Jul 2010, plus two days ahead forecasting by model in Equation (6.17).

The HCSARIMA model fitted to the 10-minutely Jul 2011 global horizontal irradiance series is given by

$$Y_t = \mu + \alpha \cos\left(\frac{2\pi}{66}\right) t + (1 + \phi_1 L + \phi_7 L^7 + \phi_9 L^9 + \phi_{12} L^{12})(1 + \theta_{10} L^{10} + \theta_{54} L^{54})\varepsilon_t + e_t. \tag{6.18}$$

Parameter estimates of the HCSARIMA model given in Equation (6.8) can be found in Appendix A, Table A.16, and the residual analysis (Ljung-Box test) results in Appendix B, Table B.16. The forecast plot for this model is given in Figure 6.34 below.

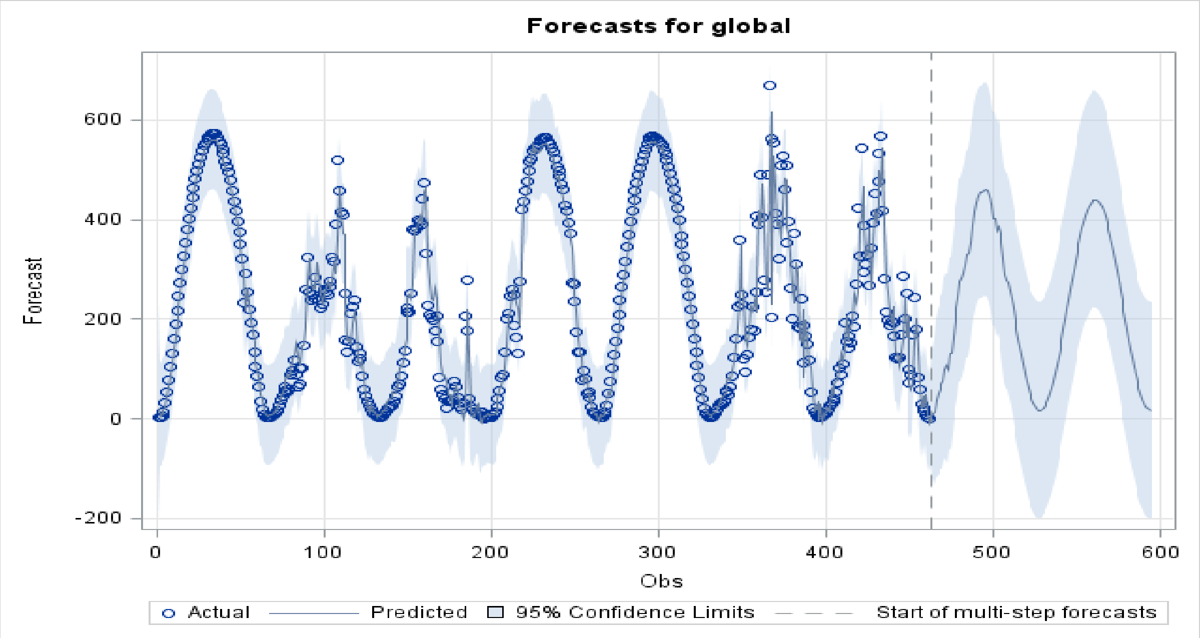

Figure 6.34: Plot of the actual versus predicted values for the 10-minutely averaged daylight global (horizontal) solar irradiance series from 3 Jul 2011 to 9 Jul 2011, plus two days ahead forecasting by model in Equation (6.18).

Similarly, as in the SARIMA models analysis, we give in Tables 6.9 to 6.12, the in-sample diagnostics and out-of-sample prediction errors for each of the HCSARIMA models fitted. The smaller the magnitude of each of these indicators, the better the fit.

Table 6.10: In-sample diagnostics for the HCSARIMA models fitted on each of the 10-minutely averaged global (horizontal) solar irradiance series.

			In-sample Model Section Diagnostics			
Model	**Season**		**AIC**	**SBC**	**R-square**	**Parsimony**
Eq. (6.11)	**Summer**	Feb 2010	-325.500	-235.775	0.986	18
Eq. (6.12)		Feb 2011	12195.330	12235.290	0.968	8
Eq. (6.13)	**Winter**	Jul 2010	252.399	290.647	0.970	8
Eq. (6.14)		Jul 2011	4955.966	4989.050	0.968	8

Table 6.11: Prediction errors for the HCSARIMA models fitted on each of the 10-minutely averaged global (horizontal) solar irradiance series.

			Model forecast accuracy measure			
Model	**Season**		**MBE**	**MPE**	**MAPE**	**RMSE**
Eq. (6.11)	**Summer**	Feb 2010	74.653	58.306	82.350	207.495
Eq. (6.12)		Feb 2011	17.134	47.301	64.775	146.817
Eq. (6.13)	**Winter**	Jul 2010	-53.844	-13.814	22.675	76.288
Eq. (6.14)		Jul 2011	-92.547	5.526	59.335	109.234

Table 6.12: In-sample diagnostics for the HCSARIMA models fitted on each of the 60-minutely averaged global (horizontal) solar irradiance series.

			In-sample Model Section Diagnostics			
Model	**Season**		**AIC**	**SBC**	**R-square**	**Parsimony**
Eq. (6.15)	**Summer**	Feb 2010	2317.351	2340.154	0.931	7
Eq. (6.16)		Feb 2011	2072.112	2091.336	0.961	6
Eq. (6.17)	**Winter**	Jul 2010	1598.668	1613.917	0.961	5
Eq. (6.18)		Jul 2011	946.258	958.413	0.883	5

Table 6.13: Prediction errors for the HCSARIMA models fitted on each of the 60-minutely averaged global (horizontal) irradiance series.

			Model forecast accuracy measure			
Model	**Season**		**MBE**	**MPE**	**MAPE**	**RMSE**
Eq. (6.15)	**Summer**	Feb 2010	185.279	457.390	457.390	239.489
Eq. (6.16)		Feb 2011	30.060	15.069	33.640	121.568
Eq. (6.17)	**Winter**	Jul 2010	-39.409	-5.273	20.159	60.397
Eq. (6.18)		Jul 2011	-89.076	-66.620	66.620	104.235

Results summary: Models and Comparisons

For ease of comparison, we present in Tables 6.14 to 6.17 below all the models for each of the years, their in-sample diagnostics used and forecast error or accuracy measures. As discussed in Chapter 4 the principle of parsimony selects the model with the least number of parameters. Clearly, the HCSARIMA models have the relatively larger AIC and SBC (BIC) values, which are due to the addition of sinusoidal predictors, compared to their respective SARIMA counterparts (see Table 6.14 and 6.16). It is also notable that the SARIMA models have wider confidence intervals for predictions indicating the higher margin of forecast error involved with this class.

Table 6.14: In-sample diagnostics for the models fitted to 2010 irradiance data.

In-sample Model Section Diagnostics						
Scale	**Date**	**Model**	**AIC**	**SBC**	**R-square**	**Parameters**
60-minutely	Feb-10	SARIMA	195.738	208.420	0.959	4
		HCSARIMA	2317.351	2340.154	0.931	7
	Jul-10	SARIMA	1490.317	1499.226	0.957	3
		HCSARIMA	1598.668	1613.917	0.961	5
10-minutely	Feb-10	SARIMA	-162.283	-123.101	0.986	8
		HCSARIMA	-325.500	-235.775	0.986	18
	Jul-10	SARIMA	400.857	419.660	0.966	4
		HCSARIMA	252.399	290.647	0.970	8

Table 6.15: Out-of-sample prediction errors compared for all the models fitted to 2010 irradiance data.

Forecast Accuracy measures						
Scale	**Date**	**Model**	**MBE**	**MPE**	**MAPE**	**RMSE**
60-minutely	Feb-10	SARIMA	185.460	101.753	109.055	286.207
		HCSARIMA	185.279	457.390	457.390	239.489
	Jul-10	SARIMA	-55.230	-37.414	37.414	64.722
		HCSARIMA	-39.409	-5.273	20.159	60.397
10-minutely	Feb-10	SARIMA	165.947	130.779	138.932	267.011
		HCSARIMA	74.653	58.306	82.350	207.495
	Jul-10	SARIMA	19.743	18.694	30.123	84.540
		HCSARIMA	-53.844	-13.814	22.675	76.288

Table 6.16: In-sample diagnostics compared for all models fitted to 2011 irradiance data.

In-sample Model Section Diagnostics						
Scale	**Date**	**Model**	**AIC**	**SBC**	**R-square**	**Parameters**
60-minutely	Feb-11	SARIMA	1973.942	1992.686	0.946	6
		HCSARIMA	2072.112	2091.336	0.961	6
	Jul-11	SARIMA	841.435	857.372	0.840	7
		HCSARIMA	946.258	958.413	0.883	5
10-minutely	Feb-11	SARIMA	11613.350	11662.500	0.958	10
		HCSARIMA	12195.330	12235.290	0.968	8
	Jul-11	SARIMA	4381.669	4409.486	0.911	7
		HCSARIMA	4955.966	4989.050	0.928	8

Table 6.17: Out-of-sample prediction errors compared for all models fitted to 2011 irradiance data.

Forecast accuracy measures						
Scale	**Date**	**Model**	**MBE**	**MPE**	**MAPE**	**RMSE**
60-minutely	Feb-11	SARIMA	39.757	10.866	50.641	143.673
		HCSARIMA	30.060	15.069	33.640	121.568
	Jul-11	SARIMA	-31.802	-26.321	63.111	45.935
		HCSARIMA	-89.076	-66.620	66.620	104.235
10-minutely	Feb-11	SARIMA	38.075	25.119	44.692	155.747
		HCSARIMA	17.134	47.301	64.775	146.817
	Jul-11	SARIMA	-94.178	-28.119	38.167	122.249
		HCSARIMA	-92.547	5.526	59.335	109.234

6.7 Long Memory (ARFIMA) Model: High frequency time series data

Box-Jenkins short memory models have been used extensively to model low frequency solar radiation series with some degree of success. However, long memory time series models known as autoregressive fractionally integrated moving average (ARFIMA) models have not been to the same extent. In this section, the efficacy of ARFIMA model to represent the underlying data generating process of the high frequency time series data is demonstrated. For testing purposes, a time series data was obtained from UKZN HC Solar Station and 20-minutely averaged values were used.

From Figure 6.36 it is evident that the long memory phenomenon is inherent in this series with the ACF plot dampening down very slowly.

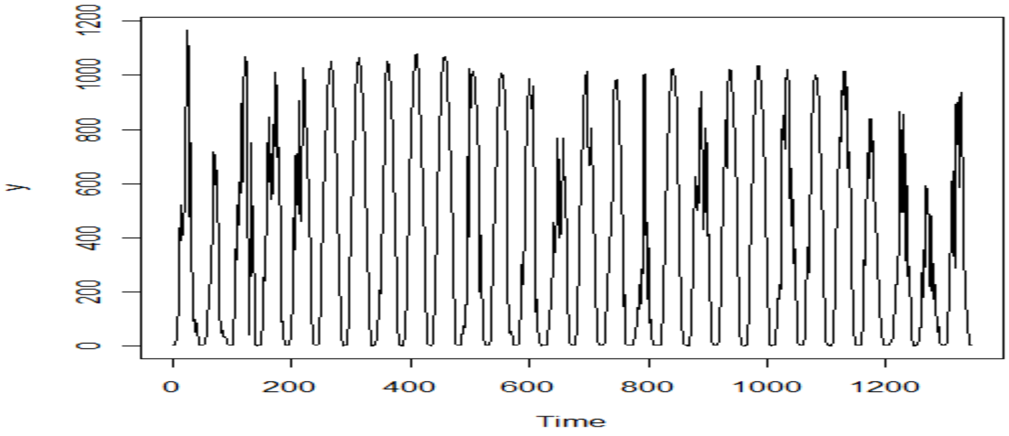

Figure 6.35: Time series plot of the 20-minutely averaged global (horizontal) solar irradiance series relating to 28 days for Feb 2010.

This ACF plot exhibits the sine-cosine waves and decays with the lag at a very low rate. The partial autocorrelation (PACF) plot dampens out. This confirms the property of a series with long range dependence (i.e. autocorrelations dying down very slowly amongst others). The long range dependence property can be captured by the long memory model. As such, the ARFIMA (1,0.40,1) model, with the long memory parameter $d = 0.4$, was fitted and estimated. The plot of the actual versus predicted values shown in Figure 6.37 indicates that the fitted model explains the underlying data generating process well. The spectrum plot of this series, with $d = 0.4$, is given in Figure 6.38. It is noted that the magnitude of the spectrum increases with the decrease in frequency and shows to diverge at frequencies near zero. Therefore, Figure 6.36 and Figure 6.38 both confirm the ARFIMA model with $d = 0.40 > 0$, based on the property “For $d > 0$ the ACF of ARFIMA time series decays very slowly and its spectrum typically diverges to infinity at frequency $\lambda = 0$, i.e. $\lim_{\lambda \to 0} f_Y(\lambda) = \infty$”, in Chapter 4.

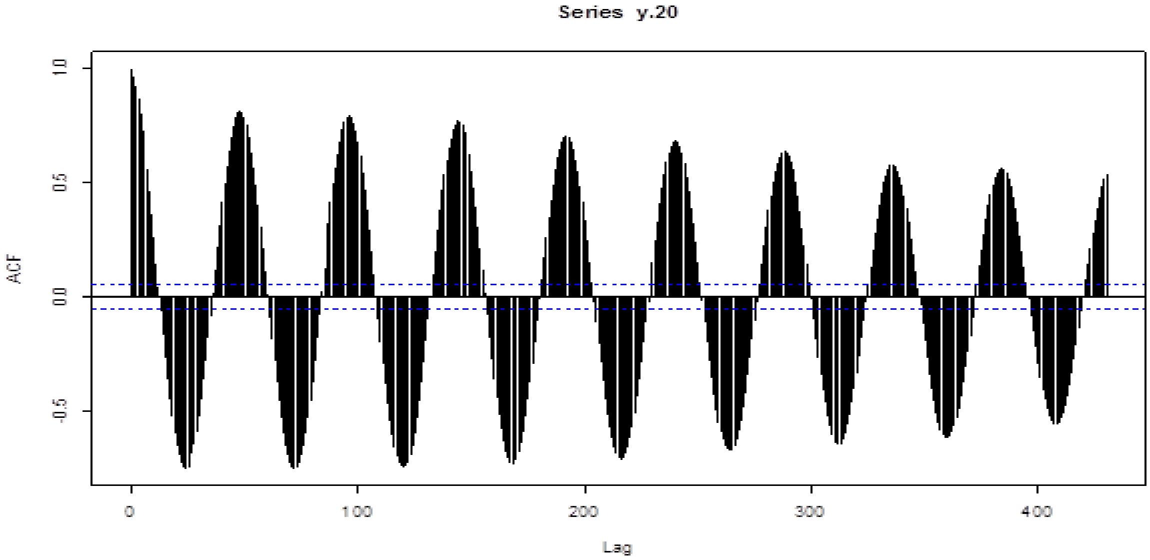

Figure 6.36: ACF plot of the 20-minutely averaged global (horizontal) solar irradiance series for Feb 2010, exhibiting the long range dependence property.

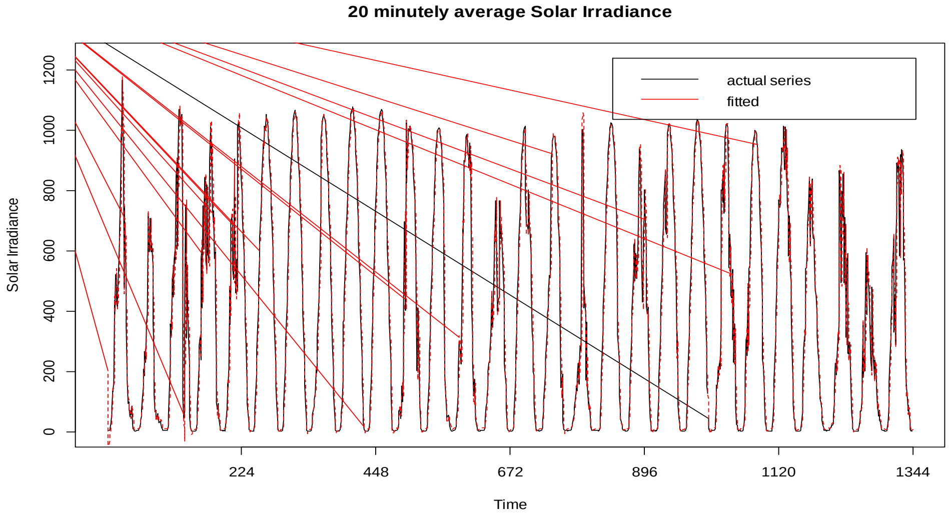

Figure 6.37: Plot of actual versus predicted values by ARFIMA (1,0.40,1) model.

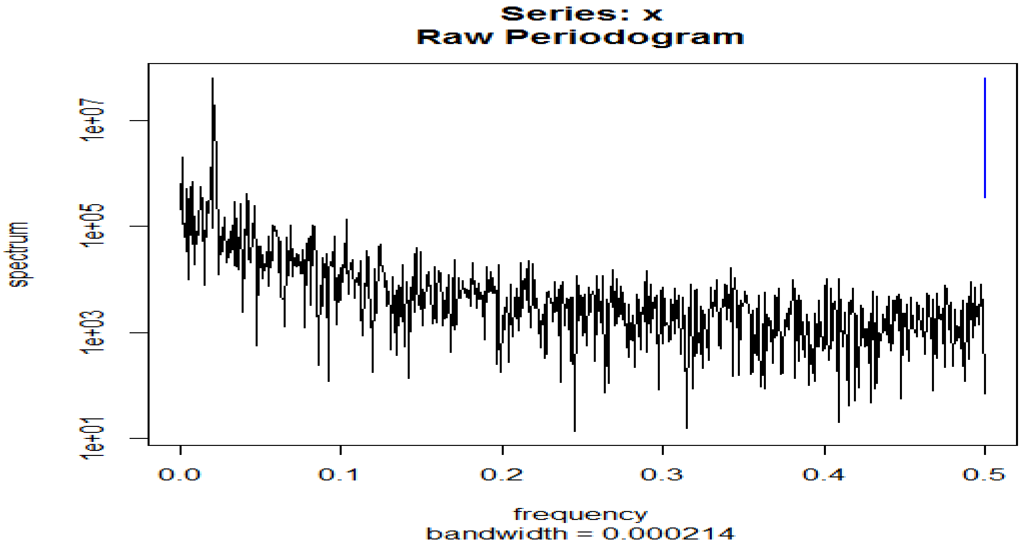

Figure 6.38: Spectrum of the 20-minutely averaged global (horizontal) solar irradiance series relating to 28 days for Feb 2010.

The ARFIMA $(1, 0.40, 1)$ process is expressed as $(1-\phi L)(1-L)^{0.4}Y_t = (1-\theta)\varepsilon_t$ or $(1-\phi L)U_t = (1-\theta)\varepsilon_t$ where $U_t = (1-L)^{0.4}Y_t \sim$ARMA $(1, 1)$, ϕ and θ are autoregressive and moving average parameters, respectively. The results of the model parameter estimation are presented in Table 6.18 below.

Table 6.18: Parameter estimation for **ARFIMA** $(\mathbf{1}, \mathbf{0.40}, \mathbf{1})$ model.

Parameter	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
d	0.40256	0.00000	Inf	<2e-16 ***
ar1	0.91541	0.00000	Inf	<2e-16 ***
ma1	0.30516	0.01197	25.48	<2e-16 ***

Chapter 7
Conclusions and Future Studies

From the results of this research study, it is concluded that the Seasonal Autoregressive Integrated Moving Average (SARIMA) and Harmonically Coupled SARIMA (HCSARIMA) classes of models both describe the underlying data generating processes of all the 10-minutely and 60-minutely averaged global horizontal irradiance time series data from UKZN HC radiometric station, with respect to various diagnostics and model predictive ability. While the two aforementioned classes of models both provided good fits for solar irradiance data series in this study, each has some distinct advantage with respect to diagnostic and prediction error analysis results (see Table 6.14 to 6.17). For example, the advantage of the HCSARIMA class of models over the SARIMA class was evident in the 2010 data with respect to forecasting accuracy (see Table 6.15).

Furthermore, the wider confidence intervals for predictions by the SARIMA class are also an indication of the higher margin of forecast error for these models. However, the clear disadvantage of the HCSARIMA class is the relatively larger AIC and SBC (BIC) values, which are due to the addition of sinusoidal predictors, compared to their respective SARIMA counterparts (see Table 6.14 and 6.16). To circumvent this problem we used a smaller number of sinusoidal predictors to model the major seasonalities. However, if the purpose of the models is only forecasting then there may be no need to restrict the number of deterministic (sinusoidal predictors) as this gives HCSARIMA models a competitive edge over SARIMA models in the prediction aspect. For 2010 data, the SARIMA models are the better class with respect to parsimony (see Table 6.15), whereas the HCSARIMA class is the best for 2011 data in the same respect.

Moreover, adding a trend component gives HCSARIMA models the competitive edge of being able to handle some aspects of second order non-stationarity, viz., presence of seasonality and trend. The search for periodicities using frequency domain techniques gives an insight into the data series that would not be detected using only time domain techniques used in the Box-Jenkins SARIMA model building methodology.

The efficacy of the autoregressive fractionally integrated moving average (ARFIMA) process to model a high frequency time series data with the long memory property was examined. From the outcome of analysis, it is tentatively concluded that the ARFIMA model is capable of capturing the long range dependence inherent in the high frequency data. Therefore, such processes are also our interest for further studies on high frequency irradiance time series data. Future work will attempt to even further improve forecast accuracy by incorporating more input parameters such as cloud cover index (see e.g. Dazhi et al., 2012) and clearness index (see e.g. Martin et al., 2010). The testing of other forecasting methods presented by literature e.g. Artificial Neural Networks (ANN) model, the Lucheroni model and the CARDS model may also form part of future work. The use of Singular Spectrum Analysis (SSA) and Multi-channel Singular Spectrum Analysis (MCSSA) is also recommended for further studies.

The models developed in this study are capable of explaining the stochastic variations of irradiance on the ground with a higher degree of accuracy than some other previously used methods, e.g. the model given in Equation (2.2). Thus, these findings are useful for generating and forecasting values of the global solar irradiance data at UKZN Howard College solar recording station with a high degree of success. The models developed in this study may help the solar system designers in setting realistic energy policies and programmes based on sound scientific principles.

Appendix A: Model Estimation in SAS

Parameter estimation for SARIMA models fitted to irradiance data

Table A.1: Parameter estimation for the SARIMA model given in Equation (6.3), fitted to 60-minutely Feb 2010 irradiance series.

Maximum Likelihood Estimation					
Parameter	**Estimate**	**Standard Error**	**t Value**	**Approx Pr > |t|**	**Lag**
MA1,1	-0.215	0.099	-2.160	0.0304	1
MA1,2	0.438	0.088	4.970	<.0001	32
AR1,1	0.487	0.095	5.130	<.0001	1
AR2,1	-0.492	0.082	-5.970	<.0001	16

Table A.2: Parameter estimation for the SARIMA model given in Equation (6.4), fitted to 60-minutely Feb 2011 irradiance series.

Maximum Likelihood Estimation					
Parameter	**Estimate**	**Standard Error**	**t Value**	**Approx Pr > |t|**	**Lag**
MA1,1	-0.590	0.073	-8.110	<.0001	1
MA1,2	-0.447	0.083	-5.420	<.0001	2
MA1,3	0.491	0.100	4.910	<.0001	28
AR1,1	0.182	0.057	3.180	0.0015	3
AR1,2	0.121	0.052	2.330	0.0198	12
AR1,3	-0.716	0.065	-11.000	<.0001	14

Table A.3: Parameter estimation for the SARIMA model given in Equation (6.5), fitted to 60-minutely Jul 2010 irradiance series.

Maximum Likelihood Estimation					
Parameter	**Estimate**	**Standard Error**	**t Value**	**Approx Pr > |t|**	**Lag**
MA1,1	0.760	0.074	10.320	<.0001	12
AR1,1	1.328	0.071	18.820	<.0001	1
AR1,2	-0.474	0.072	-6.620	<.0001	2

Table A.4: Parameter estimation for the SARIMA model given in Equation (6.6), fitted to 60-minutely Jul 2011 irradiance series.

Maximum Likelihood Estimation					
Parameter	**Estimate**	**Standard Error**	**t Value**	**Approx Pr > |t|**	**Lag**
MA1,1	-0.256	0.130	-1.970	0.0493	1
MA1,2	-0.350	0.111	-3.160	0.0016	2
AR1,1	0.199	0.078	2.550	0.0109	1
AR1,2	-0.628	0.092	-6.850	<.0001	12
AR1,3	0.132	0.058	2.270	0.0234	15
AR1,4	-0.609	0.094	-6.470	<.0001	24
AR1,5	-0.593	0.083	-7.190	<.0001	36

Table A.5: Parameter estimation for the SARIMA model given in Equation (6.7), fitted to 10-minutely Feb 2010 irradiance series.

Maximum Likelihood Estimation					
Parameter	**Estimate**	**Standard Error**	**t Value**	**Approx Pr > |t|**	**Lag**
MA1,1	-0.224	0.031	-7.150	<.0001	1
MA1,2	0.091	0.031	2.940	0.0032	3
MA1,3	-0.074	0.030	-2.480	0.0132	10
MA1,4	0.074	0.030	2.480	0.0133	13
MA2,1	0.855	0.039	21.940	<.0001	90
AR1,1	0.913	0.014	62.990	<.0001	1
AR1,2	-0.070	0.031	-2.250	0.0245	17
AR1,3	0.080	0.031	2.600	0.0092	18

Table A.6: Parameter estimation for the SARIMA model given in Equation (6.8), fitted to 10-minutely Feb 2011 irradiance series.

Maximum Likelihood Estimation					
Parameter	**Estimate**	**Standard Error**	**t Value**	**Approx Pr>|t|**	**Lag**
MA1,1	-0.422	0.135	-3.130	0.0018	2
MA1,2	0.132	0.046	2.890	0.0038	4
AR1,1	0.876	0.027	32.280	<.0001	1
AR1,2	-0.597	0.135	-4.410	<.0001	2
AR1,3	0.504	0.118	4.270	<.0001	3
AR1,4	0.094	0.028	3.320	0.0009	6
AR1,5	-0.138	0.028	-4.910	<.0001	10
AR1,6	0.097	0.027	3.600	0.0003	11
AR2,1	-0.677	0.030	-22.370	<.0001	84
AR2,2	-0.309	0.030	-10.230	<.0001	168

Table A.7: Parameter estimation for the SARIMA model given in Equation (6.9), fitted to 10-minutely Jul 2010 irradiance series.

Maximum Likelihood Estimation					
Parameter	**Estimate**	**Standard Error**	**t- Value**	**Approx Pr > |t|**	**Lag**
MA1,1	0.212	0.039	5.48	<.0001	2
AR1,1	0.751	0.035	21.21	<.0001	1
AR1,2	0.187	0.037	5.05	<.0001	2
AR2,1	-0.508	0.031	-16.34	<.0001	68

Table A.8: Parameter estimation for the SARIMA model given in Equation (6.10), fitted to 10-minutely Jul 2011 irradiance series.

Maximum Likelihood Estimation					
Parameter	**Estimate**	**Standard Error**	**t Value**	**Approx Pr > |t|**	**Lag**
AR1,1	0.863	0.026	32.93	<.0001	1
AR1,2	0.112	0.034	3.33	0.0009	7
AR1,3	-0.121	0.050	-2.42	0.0155	11
AR1,4	0.198	0.054	3.65	0.0003	12
AR1,5	-0.124	0.037	-3.31	0.0009	14
AR2,1	-0.781	0.054	-14.35	<.0001	66
AR2,2	-0.323	0.071	-4.52	<.0001	132

Parameter estimation for HCSARIMA models fitted to irradiance data

Table A.9: Parameter estimation for the HCSARIMA model given in Equation (6.11), fitted to 60-minutely Feb 2010 irradiance series, where T is the trend, COSTWO= $\cos[(2\pi/16)t]$ and SINTWO= $\cos[(2\pi/16)t]$.

Maximum Likelihood Estimation							
Parameter	**Estimate**	**Standard Error**	**t Value**	**Approx Pr > |t|**	**Lag**	**Variable**	**Shift**
MU	331.530	43.888	7.550	<.0001	0	global	0
AR1,1	0.795	0.072	11.120	<.0001	1	global	0
AR1,2	-0.214	0.072	-2.990	0.0028	2	global	0
AR2,1	0.294	0.072	4.060	<.0001	16	global	0
NUM1	0.932	0.388	2.410	0.0161	0	T	0
NUM2	-115.301	30.928	-3.730	0.0002	0	SINTWO	0
NUM3	-466.789	30.658	-15.230	<.0001	0	COSTWO	0

Table A.10: Parameter estimation for the HCSARIMA model given in Equation (6.12), fitted to 60-minutely Feb 2011 irradiance series, where $\mathrm{COSTWO} = \cos[(2\pi/14)t]$, $\mathrm{COSTWO} = \sin[(2\pi/14)t]$ and COSTHREE= $\cos[(4\pi/14)t]$.

Maximum Likelihood Estimation							
Parameter	**Estimate**	**Standard Error**	**t Value**	**Approx Pr > |t|**	**Lag**	**Variable**	**Shift**
MU	505.464	12.560	40.240	<.0001	0	global	0
AR1,1	0.668	0.055	12.130	<.0001	1	global	0
AR2,1	-0.296	0.079	-3.750	0.0002	56	global	0
NUM1	-151.068	12.156	-12.430	<.0001	0	SINTWO	0
NUM2	-456.142	12.008	-37.990	<.0001	0	COSTWO	0
NUM3	-31.363	7.617	-4.120	<.0001	0	COSTHREE	0

Table A.11: Parameter estimation for the HCSARIMA model given in Equation (6.13), fitted to 60-minutely Jul 2010 irradiance series, where COSTWO = $\cos[(2\pi/12)t]$ and SINTWO= $\sin[(2\pi/12)t]$.

Maximum Likelihood Estimation							
Parameter	**Estimate**	**Standard Error**	**t Value**	**Approx Pr > |t|**	**Lag**	**Variable**	**Shift**
MU	243.718	22.407	10.880	<.0001	0	global	0
MA1,1	-0.479	0.077	-6.190	<.0001	1	global	0
AR1,1	0.797	0.052	15.230	<.0001	1	global	0
NUM1	-68.833	12.738	-5.400	<.0001	0	SINTWO	0
NUM2	-236.917	12.626	-18.760	<.0001	0	COSTWO	0

Table A.12: Parameter estimation for the HCSARIMA model given in Equation (6.14), fitted to 60-minutely Jul 2011 irradiance series, where COSTWO= $\cos[(2\pi/12)t]$ and SINTWO= $\sin[(2\pi/12)t]$.

Maximum Likelihood Estimation							
Parameter	**Estimate**	**Standard Error**	**t Value**	**Approx Pr > |t|**	**Lag**	**Variable**	**Shift**
MU	214.435	34.672	6.180	<.0001	0	global	0
AR1,1	0.745	0.073	10.180	<.0001	1	global	0
AR2,1	0.277	0.118	2.340	0.0193	15	global	0
NUM1	-62.940	18.870	-3.340	0.0009	0	SINTWO	0
NUM2	-210.121	18.567	-11.320	<.0001	0	COSTWO	0

Table A.13: Parameter estimation for the HCSARIMA model given in Equation (6.15), fitted to 10-minutely Feb 2010 irradiance series, where $\mathrm{COSTWO} = \cos[(2\pi/90)t]$, $\mathrm{SINTWO} = \sin[(2\pi/90)t$, $\mathrm{COSTHREE} = \cos[(4\pi/90)t]$ and $\mathrm{SINTHREE} = \sin[(4\pi/90)t]$.

Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable	Shift
MU	5.180	0.139	37.240	<.0001	0	globallog	0
MA1,1	-0.983	0.089	-11.000	<.0001	1	globallog	0
MA1,2	-0.295	0.034	-8.780	<.0001	2	globallog	0
MA2,1	-0.154	0.031	-4.940	<.0001	88	globallog	0
MA2,2	-0.120	0.031	-3.850	0.0001	89	globallog	0
AR1,1	0.215	0.090	2.390	0.0168	1	globallog	0
AR1,2	0.607	0.083	7.270	<.0001	2	globallog	0
AR1,3	-0.103	0.021	-4.950	<.0001	11	globallog	0
AR1,4	0.087	0.021	4.110	<.0001	21	globallog	0
AR2,1	0.070	0.030	2.300	0.0216	55	globallog	0
AR2,2	-0.076	0.030	-2.510	0.0122	81	globallog	0
AR2,3	0.163	0.030	5.370	<.0001	90	globallog	0
AR2,4	0.137	0.030	4.510	<.0001	92	globallog	0
AR2,5	0.087	0.030	2.890	0.0039	93	globallog	0
NUM1	-0.554	0.121	-4.590	<.0001	0	SINTWO	0
NUM2	-2.186	0.120	-18.30	<.0001	0	COSTWO	0
NUM3	-0.524	0.114	-4.590	<.0001	0	SINTHREE	0
NUM4	-0.829	0.113	-7.310	<.0001	0	COSTHREE	0

Table A.14: Parameter estimation for the HCSARIMA model given in Equation (6.16), fitted to 10-minutely Feb 2011 irradiance series, where COSTWO= $\cos[(2\pi/84)t]$ and SINTWO= $\sin[(2\pi/84)t]$.

Maximum Likelihood Estimation							
Parameter	**Estimate**	**Standard Error**	**t Value**	**Approx Pr > |t|**	**Lag**	**Variable**	**Shift**
MU	498.676	14.426	34.570	<.0001	0	global	0
AR1,1	0.858	0.030	28.410	<.0001	1	global	0
AR1,2	-0.145	0.039	-3.680	0.0002	2	global	0
AR1,3	0.076	0.032	2.390	0.0167	3	global	0
AR1,4	0.145	0.031	4.620	<.0001	6	global	0
AR1,5	-0.067	0.030	-2.230	0.026	7	global	0
NUM1	-48.184	16.437	-2.930	0.0034	0	SINTWO	0
NUM2	-478.809	16.261	-29.440	<.0001	0	COSTWO	0

Table A.15: Parameter estimation for the HCSARIMA model given in Equation (6.17), fitted to 10-minutely Jul 2010 irradiance series, where COSTWO= $\cos[(2\pi/68)t]$ and COSTHREE= $\cos[(4\pi/68)t]$.

Maximum Likelihood Estimation							
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable	Shift
MU	4.836	0.286	16.920	<.0001	0	globallog	0
MA1,1	0.968	0.008	123.910	<.0001	68	globallog	0
AR1,1	0.806	0.025	32.170	<.0001	1	globallog	0
AR1,2	0.057	0.025	2.260	0.0236	3	globallog	0
AR1,3	0.060	0.017	3.470	0.0005	17	globallog	0
AR2,1	0.999	0.000	3761.170	<.0001	68	globallog	0
SCALE1	-1.844	0.151	-12.210	<.0001	0	COSTWO	0
SCALE2	-0.957	0.110	-8.660	<.0001	0	COSTHREE	0

Table A.16: Parameter estimation for the HCSARIMA model given in Equation (6.18), fitted to 10-minutely Jul 2011 irradiance series, where COSTWO= $\cos[(2\pi/66)t]$.

Maximum Likelihood Estimation							
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable	Shift
MU	228.483	25.364	9.010	<.0001	0	global	0
MA1,1	-0.135	0.055	-2.460	0.014	10	global	0
MA1,2	0.150	0.052	2.900	0.0037	54	global	0
AR1,1	0.831	0.028	29.970	<.0001	1	global	0
AR1,2	0.130	0.038	3.430	0.0006	7	global	0
AR1,3	-0.143	0.045	-3.160	0.0016	9	global	0
AR1,4	0.091	0.036	2.530	0.0114	12	global	0
NUM1	-211.674	18.988	-11.150	<.0001	0	COSTWO	0

Appendix B: Model Residual Analysis in SAS

Residual analysis for SARIMA models fitted to irradiance data

Table B.1: Residual analysis for the SARIMA model given in Equation (6.3).

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	4.270	2	0.118	0.009	-0.006	-0.078	0.111	0.028	0.064
12	11.820	8	0.159	0.028	-0.055	0.169	0.039	-0.038	-0.070
18	20.150	14	0.126	0.080	0.089	0.121	0.013	-0.116	-0.017
24	26.820	20	0.140	0.151	-0.031	-0.047	0.001	-0.005	0.084
30	31.500	26	0.210	0.066	-0.054	-0.070	0.088	0.007	0.049
36	35.050	32	0.326	0.028	0.061	-0.017	0.030	0.080	-0.062
42	35.260	38	0.597	-0.004	-0.008	-0.013	0.000	0.025	0.006
48	38.970	44	0.687	-0.003	0.084	-0.000	0.002	0.089	0.019

Table B.2: Residual analysis for the SARIMA model given in Equation (6.4).

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	.	0	.	0.112	-0.026	0.032	0.043	-0.124	-0.097
12	8.280	6	0.218	0.003	-0.012	0.018	-0.014	0.075	-0.011
18	12.620	12	0.397	-0.074	-0.053	-0.016	0.072	0.086	0.045
24	18.990	18	0.393	-0.008	-0.084	-0.048	0.055	0.090	-0.109
30	29.440	24	0.204	-0.080	0.007	-0.080	-0.120	-0.155	0.012
36	34.590	30	0.258	-0.063	0.056	0.067	0.035	0.086	0.064
42	39.000	36	0.336	-0.017	-0.054	0.005	-0.064	0.062	-0.093
48	42.470	42	0.451	-0.006	-0.030	0.036	-0.088	-0.001	0.069

Table B.3: Residual analysis for the SARIMA model given in Equation (6.5).

Autocorrelation Check of Residuals									
To Lag	**Chi-Square**	**DF**	**Pr > ChiSq**	**Autocorrelations**					
6	3.770	3	0.287	0.014	-0.063	0.130	-0.058	-0.010	-0.029
12	14.070	9	0.120	-0.048	-0.047	0.190	-0.077	-0.008	0.138
18	16.180	15	0.370	-0.093	-0.046	-0.004	-0.037	-0.010	0.030
24	18.090	21	0.643	-0.008	-0.065	0.031	0.019	-0.040	-0.062
30	19.900	27	0.835	0.046	-0.012	0.029	-0.040	-0.060	0.042
36	23.410	33	0.891	0.008	-0.009	0.066	-0.055	-0.047	0.092
42	25.010	39	0.960	-0.039	-0.070	0.015	-0.026	-0.022	0.015
48	30.440	45	0.952	0.037	0.023	-0.046	-0.127	0.017	0.070

Table B.4: Residual analysis for the SARIMA model given in Equation (6.6).

Autocorrelation Check of Residuals									
To Lag	**Chi-Square**	**DF**	**Pr > ChiSq**	**Autocorrelations**					
6	.	0	.	0.032	0.000	0.109	-0.02	0.146	0.060
12	8.170	5	0.147	-0.156	0.122	0.049	-0.014	-0.013	-0.138
18	12.320	11	0.340	0.121	0.094	0.044	0.011	-0.057	-0.124
24	18.570	17	0.354	0.145	-0.135	-0.105	0.05	-0.068	-0.055
30	25.970	23	0.302	0.062	-0.178	0.047	0.072	-0.122	0.065
36	27.150	29	0.563	-0.043	-0.006	0.068	0.000	-0.004	0.045
42	32.230	35	0.603	-0.013	0.106	-0.063	0.038	0.117	-0.019
48	41.960	41	0.429	-0.018	0.145	-0.030	0.033	-0.036	-0.150

Table B.5: Residual analysis for the SARIMA model given in Equation (6.7).

Autocorrelation Check of Residuals									
To Lag	**Chi-Square**	**DF**	**Pr > ChiSq**	**Autocorrelations**					
6	.	0	.	0.002	0.017	0.005	0.012	-0.047	0.045
12	6.160	4	0.187	-0.009	-0.013	0.006	-0.012	-0.029	0.013
18	8.630	10	0.568	-0.003	-0.026	0.031	-0.025	-0.001	-0.014
24	15.800	16	0.467	-0.031	0.024	0.013	-0.010	0.072	0.007
30	26.030	22	0.251	-0.019	0.003	-0.075	-0.032	-0.003	0.054
36	34.850	28	0.174	-0.057	0.000	-0.021	0.026	0.014	0.063
42	40.460	34	0.207	-0.025	0.030	0.024	-0.020	0.051	-0.016
48	53.320	40	0.077	-0.084	-0.004	-0.013	-0.034	0.053	0.036

Table B.6: Residual analysis for the SARIMA model given in Equation (6.8).

Autocorrelation Check of Residuals									
To Lag	**Chi-Square**	**DF**	**Pr> ChiSq**	**------------------Autocorrelations------------**					
6	.	0	.	0.018	0.001	0.002	0.000	-0.040	-0.010
12	3.040	2	0.219	0.009	-0.013	0.022	0.005	0.004	0.014
18	9.730	8	0.285	-0.015	0.036	-0.033	0.016	-0.060	-0.008
24	11.470	14	0.649	0.021	0.016	0.027	0.016	0.002	-0.002
30	15.830	20	0.727	0.015	-0.025	0.043	0.003	0.012	-0.037
36	23.660	26	0.595	-0.042	-0.034	-0.029	0.012	-0.034	0.049
42	25.680	32	0.778	-0.012	-0.035	-0.010	0.007	-0.017	-0.010
48	29.910	38	0.823	-0.007	0.044	0.037	-0.001	0.025	-0.005

Table B.7: Residual analysis for the SARIMA model given in Equation (6.9).

Autocorrelation Check of Residuals									
To Lag	**Chi-Square**	**DF**	**Pr > ChiSq**	**Autocorrelations**					
6	5.710	2	0.058	0.005	0.007	-0.056	-0.013	-0.031	-0.051
12	8.390	8	0.396	0.002	0.026	0.035	0.004	0.028	0.022
18	9.730	14	0.781	-0.003	-0.005	0.002	0.012	0.036	-0.010
24	12.490	20	0.898	-0.042	0.026	-0.009	-0.013	0.020	-0.017
30	15.600	26	0.945	-0.015	0.002	0.050	0.010	0.029	0.003
36	20.210	32	0.948	0.063	-0.013	0.004	0.013	0.027	0.019
42	20.650	38	0.990	-0.007	0.005	-0.019	0.006	0.000	-0.006
48	23.390	44	0.995	0.030	0.020	0.018	-0.036	0.013	-0.009

Table B.8: Residual analysis for the SARIMA model given in Equation (6.10).

Autocorrelation Check of Residuals									
To Lag	**Chi-Square**	**DF**	**Pr > ChiSq**	**Autocorrelations**					
6	.	0	.	-0.035	-0.003	0.063	-0.022	0.057	-0.015
12	9.480	5	0.091	0.025	0.027	-0.102	0.051	0.004	0.003
18	11.150	11	0.431	-0.021	-0.034	0.033	0.008	-0.019	-0.032
24	13.860	17	0.677	0.038	0.005	0.018	-0.043	-0.044	-0.030
30	20.330	23	0.622	0.058	0.069	0.057	0.021	0.021	-0.054
36	21.190	29	0.852	-0.014	0.008	0.009	0.035	0.009	-0.019
42	24.040	35	0.919	0.016	-0.045	-0.046	0.022	-0.040	0.001
48	28.130	41	0.937	0.011	-0.059	0.025	0.023	-0.059	0.030

Residual analysis for HCSARIMA models fitted to irradiance data

Table B.9: Residual analysis for the HCSARIMA model given in Equation (6.11).

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr> ChiSq	--------------------Autocorrelations--------------------					
6	4.520	3	0.211	0.027	-0.069	0.019	0.095	0.077	0.047
12	6.210	9	0.718	-0.033	-0.026	0.057	-0.009	-0.054	-0.018
18	9.310	15	0.861	-0.020	0.109	0.021	-0.006	-0.013	0.042
24	11.550	21	0.951	-0.026	-0.092	-0.035	0.001	0.007	-0.007
30	16.690	27	0.939	-0.033	-0.040	-0.017	-0.008	0.118	-0.074
36	23.590	33	0.886	-0.022	-0.027	0.082	-0.140	0.020	0.037
42	28.100	39	0.902	0.028	-0.127	0.002	0.023	-0.019	-0.028
48	34.020	45	0.884	0.060	0.051	0.099	0.070	0.047	-0.009

Table B.10: Residual analysis for the HCSARIMA model given in Equation (6.12).

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	--------------------Autocorrelations----------------					
6	5.070	4	0.280	0.032	0.024	-0.014	-0.001	-0.090	-0.130
12	11.020	10	0.356	-0.020	-0.017	0.045	-0.044	0.115	0.111
18	18.980	16	0.270	-0.152	-0.041	-0.108	-0.009	-0.057	-0.003
24	23.520	22	0.373	-0.004	-0.078	-0.022	-0.006	0.099	-0.072
30	30.240	28	0.352	-0.048	-0.093	-0.099	-0.044	-0.002	0.091
36	38.970	34	0.256	-0.072	0.060	0.136	-0.006	0.081	0.070
42	44.350	40	0.294	-0.061	-0.057	0.044	-0.057	0.061	-0.083
48	48.730	46	0.364	-0.001	-0.048	0.053	-0.071	0.085	0.021

Table B.11: Residual analysis for the HCSARIMA model given in Equation (6.13).

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr> ChiSq	--------------------Autocorrelations--------------------					
6	4.060	4	0.398	0.013	0.073	0.077	-0.083	-0.029	-0.077
12	10.670	10	0.384	-0.012	-0.065	0.108	-0.079	-0.086	0.097
18	17.540	16	0.351	-0.174	-0.066	-0.042	-0.005	0.019	0.053
24	19.320	22	0.625	0.029	-0.024	0.005	-0.030	-0.081	-0.027
30	21.130	28	0.820	-0.059	-0.052	0.023	-0.024	-0.012	0.046
36	23.880	34	0.902	0.043	0.031	-0.002	-0.018	-0.065	0.079
42	27.560	40	0.932	-0.110	-0.066	0.008	-0.008	0.020	0.025
48	31.550	46	0.948	0.045	0.023	-0.083	-0.088	-0.016	0.024

Table B.12: Residual analysis for the HCSARIMA model given in Equation (6.14).

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	--------------------Autocorrelations---------------- ----					
6	4.430	4	0.351	-0.055	0.094	-0.011	-0.056	0.181	-0.032
12	7.290	10	0.698	-0.035	-0.036	0.045	-0.103	-0.058	-0.102
18	11.970	16	0.746	0.010	0.067	0.026	0.137	0.038	-0.134
24	23.920	22	0.351	0.101	-0.183	-0.018	-0.071	-0.176	-0.152
30	29.080	28	0.409	-0.031	-0.132	-0.056	-0.011	-0.129	0.047
36	32.420	34	0.545	-0.042	0.074	0.081	0.038	0.019	-0.088
42	36.860	40	0.613	-0.098	0.087	-0.084	0.037	0.031	-0.035
48	41.680	46	0.654	0.045	0.106	-0.016	0.011	0.040	0.101

Table B.13: Residual analysis for the HCSARIMA model given in Equation (6.15).

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	--------------------Autocorrelations----------------					
6	.	0	.	0.000	0.005	0.000	-0.016	-0.022	-0.001
12	.	0	.	0.000	0.016	0.003	0.013	0.023	0.025
18	10.530	5	0.062	-0.034	-0.041	0.029	-0.027	-0.049	0.022
24	15.110	11	0.177	0.007	0.051	-0.012	-0.001	0.037	-0.004
30	24.120	17	0.116	-0.008	0.043	-0.048	-0.017	0.027	0.053
36	35.830	23	0.043	-0.076	-0.010	-0.007	0.048	0.016	0.045
42	39.460	29	0.093	-0.017	0.014	-0.006	-0.028	0.036	-0.024
48	51.660	35	0.035	-0.067	-0.002	-0.007	-0.047	0.056	0.030

Table B.14: Residual analysis for the HCSARIMA model given in Equation (6.16).

Autocorrelation Checks of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	--------------------Autocorrelations--------------------					
6	0.540	1	0.463	-0.001	0.001	0.020	-0.005	-0.004	-0.008
12	11.620	7	0.114	0.038	-0.019	0.014	-0.075	0.010	0.049
18	21.250	13	0.068	-0.008	0.019	-0.025	0.043	-0.075	-0.012
24	22.480	19	0.261	0.012	0.016	0.016	-0.005	-0.007	-0.019
30	26.010	25	0.407	-0.004	-0.012	0.036	0.014	0.006	-0.038
36	30.620	31	0.486	-0.026	-0.021	-0.029	0.009	-0.023	0.039
42	32.740	37	0.669	-0.006	-0.038	0.000	0.007	0.013	0.013
48	36.680	43	0.741	-0.003	0.035	0.022	0.011	0.037	-0.015

Table B.15: Residual analysis for the HCSARIMA model given in Equation (6.17).

Autocorrelations of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	--------------------Autocorrelations--------------------					
6	3.460	1	0.063	-0.006	0.014	0.014	0.019	0.001	-0.056
12	5.990	7	0.541	-0.028	-0.016	-0.003	-0.031	0.029	0.007
18	11.180	13	0.596	0.005	0.023	0.032	0.046	0.027	-0.036
24	14.840	19	0.733	-0.038	0.045	-0.007	-0.014	0.018	-0.002
30	15.750	25	0.922	-0.003	-0.004	0.024	0.009	0.016	0.009
36	18.070	31	0.969	0.037	-0.019	-0.022	-0.017	-0.007	-0.002
42	20.210	37	0.989	-0.015	-0.022	-0.036	0.015	0.003	-0.010
48	23.440	43	0.993	0.028	0.027	0.036	0.004	0.021	0.012

Table B.16: Residual analysis for the HCSARIMA model given in Equation (6.18).

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	--------------------Autocorrelations--------------------					
6	.	0	.	-0.030	-0.031	0.104	-0.029	0.038	-0.004
12	10.080	6	0.121	-0.001	0.028	0.034	0.014	-0.027	0.060
18	15.110	12	0.235	0.046	-0.069	-0.026	-0.005	-0.009	-0.053
24	17.840	18	0.466	0.017	0.005	0.019	-0.039	-0.033	-0.048
30	24.970	24	0.408	0.052	0.063	-0.003	0.032	0.031	-0.076
36	28.000	30	0.570	-0.015	0.018	0.004	0.052	-0.017	-0.050
42	32.630	36	0.630	0.036	-0.023	-0.065	0.014	-0.054	0.006
48	37.400	42	0.673	0.020	-0.058	0.014	0.042	-0.059	-0.002

Appendix C: Sample SAS and R Programs for HCSARIMA and ARFIMA models

global = original variable
globallog = log-transformation of the original variable

Typical SAS Program for the HCSARIMA model given in Equation 6.15 fitted to sixty-minutely averaged time series data for July 2011

```
title "60-minute averaged global solar irradiance for Jul 2011";
data solar;
input datetime : datetime15. global @@;
format datetime datetime12.;
  hour = hour( datetime );
  date = datepart( datetime );
  year = year( date );
  month = month( date );
  day = day( date );
globallog = log(global);
T  = _n_;
PI=ARCOS(-1);
TWO=(PI*T)/6;
SINTWO=SIN(TWO);
COSTWO=COS(TWO);
label global = 'global;
       globallog = 'log transformed global'
        T  = 'Time, 60 minutely';
datalines;
3Jul2011:06:00        4.740425167
3Jul2011:07:00        68.77166
3Jul2011:08:00        230.7256
.        .
.        .
.        .
.        .
.        .
9Jul2011:15:00        153.9708
9Jul2011:16:00        65.08139
9Jul2011:17:00        5.071646

;
proc sgplot data=solar;
scatter x=T y=global;
```

```
series x=T y=global;
run;

title "60-minute averaged log transformed global solar irradiance for July 2011";
proc sgplot data=solar;
scatter x=T y=globallog;
series x=T y=globallog;
run;

title "60-minute global solar irradiance for July 2011";
proc spectra data=solar out=b p s adjmean whitetest;
    var global;
    weights 1 1 1 1 1;
  run;

  proc print data=b;
  run;


proc sgplot data=b;
label p_01 = 'Periodogram of global irradiance';
scatter x=period y=p_01;
series x=period y=p_01;
run;

title "60-minute averaged log transformed global solar irradiance for July 2011";
proc spectra data=solar out=b p s adjmean whitetest;
    var globallog;
    weights 1 1 1 1 1;
  run;

  proc print data=b;
  run;


proc sgplot data=b;
label p_01 = 'Periodogram of log transformed global irradiance';
scatter x=period y=p_01;
series x=period y=p_01;
run;


proc arima data=solar plots=all;
identify var=global crosscorr=(T SINTWO COSTWO) nlag=124 esacf stationarity=(adf=(0 1 2 3
4 5));
run;
```

```
estimate input=(T SINTWO COSTWO ) ml; run;


/*FINAL*/
estimate input=(SINTWO COSTWO ) p=(1)(15) ml;
forecast id=datetime interval=minute60 lead=24  printall out=w;
/*forecast lead=36 out=predict printall out=e;*/
run;

title "ARIMA forecasts of the log of global irradiance";
data z;
     set w;
      global = global;
     Dl95 = l95;
     Du95 = u95;
    HCSARIMA = forecast;

  run;

title "Forecasts of global irradiance";
  proc sgplot data=z;
where datetime >= '02Jul2011:04:00'dt;
     band Upper=Du95 Lower=Dl95 x=datetime / transparency=.75 legendlabel="HCSARIMA
95% Confidence Limits" fillattrs=(color=red);
     /*/ LegendLabel="95% Confidence Limits";*/
  scatter x=datetime y=global;
  series x=datetime y=HCSARIMA / markers
 markerattrs=(color=red) lineattrs=(color=red) LegendLabel="Forecast for global";
/*/ LegendLabel="Forecast for global" ;*/
refline '9Jul2011:19:00'dt / axis=x;
run;


proc arima data=solar plots=all;
identify var=global crosscorr=(T SINTWO COSTWO) nlag=124 esacf stationarity=(adf=(0 1 2 3
4 5));
run;

estimate input=(T SINTWO COSTWO ) ml; run;


/*FINAL*/
estimate input=(SINTWO COSTWO ) p=(1)(15) ml;
forecast id=datetime interval=minute60 lead=24  printall out=w;
run;
title "ARIMA forecasts of the log of global irradiance";
```

```
data z;
      set w;
       global = global;
      Dl95 = l95;
      Du95 = u95;
     HCSARIMA =  forecast ;

   run;
data exe;
   merge solar (drop=globallog  SINTWO COSTWO)
                  z;
run;

proc print data=exe;
   /*title 'Acting Class Exercise Schedule';*/
run;

title "Series from 3 to 9 July, 2011 plus 2 days ahead forecasts";
   proc sgplot data=exe;
where T >= 13;
      band Upper=Du95 Lower=Dl95 x=T / transparency=.75 legendlabel="HCSARIMA 95%
Confidence Limits" fillattrs=();
      /*/ LegendLabel="95% Confidence Limits";*/
   scatter x=T y=global;
   series x=T y=HCSARIMA / markers
  markerattrs=() lineattrs=() LegendLabel="Forecast for global";
/*/ LegendLabel="Forecast for global" ;*/
 refline 84 / axis=x;
run;

data exer;
   merge solar (drop=global globallog)
                 f z;
run;

proc print data=exer;
   /*title 'Acting Class Exercise Schedule';*/
run;
```

Typical R Program for the ARFIMA model fitted to twenty-minutely averaged time series data

```
 library(arfima)
library(forecast)

y.20=scan("I:\\Feb20min.txt")

fit.20 = armaFit( ~ arfima(1, 1) , data = y.20)
 fit.20
summary(fit.20)


fit.21=arfima(y.20,max.p=1,max.q=1)
fit.21
plot(forecast(fit.21,h=30))

plot(fit.21$fitted)
 plot(forecast(fit.21,h=150))


plot(1:length(y.20),y.20,type = "l", ylim=c(0,1240),xaxt="n", xlab="Time", ylab="Solar
Irradiance" )
title("20 minutely average Solar Irradiance")

axis(side=1,at=c(,,,,,))

lines(1:length(y.20), fit.21$fitted, col=2,lty = 2)

leg <- c("actual series","fitted")
     legend(length(y.20)-500, 1239, legend=leg, lty=1, col=1:2)


spectrum(y.20, spans = NULL)
spectrum(y.20, spans = c(3,5))


#DAILY
y.d=scan("i:\\daily10.txt")
fit.d=arima(y.d,order = c(0, 1, 2))
 summary(fit.d)

yd.fitted=y.d-fit.d$resid

plot(1:length(y.d),y.d,type = "l", ylim=c(0,740),xaxt="n", xlab="Time", ylab="Solar Irradiance")
title("Daily average Solar Irradiance")
```

```
axis(side=1,at=c(55,110,165,220,275,330))

lines(1:length(y.d), yd.fitted, col=2,lty = 2)

leg <- c("actual series","fitted")
    legend(length(y.d)-220, 739, legend=leg, lty=1, col=1:2)
```

References

1. Akaike H. (1983). Information Measures and Model Selection. *Bulletin of the International Statistical Institute*, 50, 277-290.
2. Alam S., Kaushik S.C. and Garg S.N. (2006). Computation of beam solar radiation at normal incidence using artificial neural network. *Renewable Energy*, 31, 1483–1491.
3. Almorox J., Hontoria C. and Benito M. (2005). Statistical validation of daylength definitions for estimation of global solar radiation. *Conversion and Management*, 45, 1465 – 1471.
4. Angstrom A. (1924). Solar and terrestrial radiation. Q.J. *Roy. Met. Soc.*, 4,121–126.
5. Arun P., Banerjee R. and Bandyopadhyay S. (2006). Sizing curve for design of isolated systems. *Advances in Energy Research.*
6. Ashley R.A. and Patterson D.M. (2010). Apparent long memory in time series as an artifact of a time-varying mean: Considering alternatives to the fractionally integrated model. *Macroeconomics Dynamics*, 14, 59-87.
7. Badescu V. (2008). *Modelling Solar Radiation at the Earth's Surface*. Springer, Berlin Heidelberg.
8. Baldocchi D. (2012). Lecture 7, Solar Radiation, Part 3, Earth-Sun Geometry. *Biometeorology*, ESPM 129.
9. Barbaro S., Cannata G., Coppolino S., Leone C. and Sinagra E. (1981). Correlation between relative sunshine and state of the sky. *Solar Energy*, 26, 537–550.
10. Bartlett M.S. (1963). The spectral analysis of point processes. *J. R. Statist. Soc. B*, 25, 264-296.
11. Beran G.W. and Terrin N. (1994). Estimation of the long-memory parameter based on a multivariate central limit theorem. *Journal of Time Series Analysis* 15,269-278.
12. Bermudez J.D., Segura J.V. and Vercher E. (2005). Holt-Winters forecasting: an alternative formulation applied to UK air passenger data. *Centro de Investigacion Operativa.*
13. Boutahar M. and Khalfaoui R. (2011). Estimation of the long memory parameter in non-stationary models. *GREQAM*.
14. Box G.E. and Cox D.R. (1964). An analysis of transformations. *J. Roy. Statist. Soc. Ser. B* 26:211–252.
15. Box G.E.P and Jenkins G.M. (1976). Time Series Analysis: Forecasting and Control. *Operational Research Quarterly*, 22, 199-201.
16. Brietzke E.H.M., Lopes S.R.C. and Bisognin C. (2005). A Clossed Formular for the Durbin-Levinson's Algorithm in Seasonal Fractionally Integrated Processes. *Mathematical and Computer Modelling*, 42, 1191-1206.

17. Brown R.G. (1956). Exponential smoothing for predicting demand. Tenth national meeting of the Operation Research Society of America, San Francisco.
18. Chen W., Mahlke S., Warter N., Anik S. and Hwu, W. (1994). Profile assisted instruction scheduling. *Int. J. of Parallel Programming* 22, 151-181.
19. Chiawa M.A., Asare B.K. and Audu B. (2010). Short and long memory time series models of relative humidity of Jos Metropolis. *Res. J. Math. Stat.* 2, 23-31.
20. Craggs C., Conway E. and Pearsall N.M. (1999). Stochastic modelling of solar irradiance on horizontal and vertical planes at a northerly location. *Renewable Energy*, 18, 445-463.
21. Davis H.T. (1941). *The Analysis of Economic Time Series*. Indiana: The Principia Press, Inc. Bloomington, 1941.
22. DeLurgio S.A. (1998). Forecasting Principles and Applications. *McGraw Hill, N.Y.*
23. Dogniaux R. and Lemoine M. (1983). Classification of radiation sites in terms of different indices of atmospheric transparency. *Solar Energy Research and Development in the European Community Series F*, 2, 94–107.
24. Egido M. and Lorenzo E. (1992). The sizing of stand alone PV-system: a review and a proposed new method. *Solar Energy Materials and Solar Cells*, 26, 51-69.
25. Engle R. (2001). GARCH 101: The Use of ARCH/GARCH Models in Applied Econometrics. *Journal of Economic Perspectives* 15, 157–168.
26. Froehlich C. and Brusa R.W. (1981). Solar radiation and its variation in time. *Solar Physics*, 74, 209-215.
27. Froehlich C. and Brusa R.W. (1981). Solar Radiation and its Variation in Time. *Sol Phys*., 74, 209–215.
28. Fuller W.A. (1976). *Introduction to Statistical Time Series*. New York: John Wiley & Sons.
29. Geweke J. and Porter-Hudak S. (1983). The estimation and application of long memory time series models. *J. Time Ser. Anal.* 4, 221±37.
30. Glover J. and McCulloch J.S. (1958). The empirical relation between solar radiation and hours of sunshine. *Quart J.R. Met. Soc.* 84:172.
31. Gopinathan K.K. (1988). A general formula for computing the coefficients of the correlations connecting global solar radiation to sunshine duration. *Solar Energy*, 41:499–502.
32. Granger C.W.J. (1980). Long memory relationships and the aggregation of dynamic models. *Journal of Econometrics*, 14, 227-238.
33. Granger C.W.J. and Joyeux R. (1980). An introduction to long-memory time series and fractional differencing. *Journal of Time Series Analysis 1, forthcoming*.

34. Gueymard C.A., Myers D. and Emery K. (2002). Proposed Reference Irradiance Spectra for Solar Energy Systems Testing. *Solar Energy*, 73, 443-467.

35. Gupta P. and Srinivasan R. (2011). Missing Data Prediction and Forecasting for Water Quantity Data. *International Conference on Modelling, Simulation and Control*.

36. Haslett J. and Raftery E. (1989). State-Space Modelling with Long Memory Dependence. Assessing Ireland's Wind Power Resource. *Applied Statistics*, 38, 1 – 50.

37. Holt C.C. (1957). Forecasting seasonals and trends by exponentially weighted moving averages, *ONR Memorandum*, 52. Pittsburgh, PA: Carnegie Institute of Technology. Available from the Engineering Library, University of Texas at Austin.

38. Hosking J.R.A. (1981). Fractional differencing. *Biometrika*, 68, 165-176.

39. http:// www.kzngreengrowth.com

40. http:// www.newport.com/Introduction-to-Solar-Radiation

41. http://gradrad.ukzn.ac.za, GRADRAD: The Greater Durban Radiometric Network

42. http://www.elsivier.com/locate/solener

43. http://www.powerfromthesun.net/chapter2/Chapter2.htm. The Sun's Energy-Power From The Sun

44. Huang J., Korolkiewicz M., Agrawal M. and Boland J. (2011). Forecasting solar radiation on short time scales using a coupled autoregressive and dynamical system (CARDS) model. *Solar Energy*, 87, 136-149.

45. Huo J., Cox C.D., Seaver W.L., Robinson R.B. and Jiang Y. (2010). Application of Two-Directional Time Series Models to Replace Missing Data. *Journal of Environmental Engineering*, 136, 435-443.

46. Hurst H.E. (1951). Long term storage capacity of reservoirs, Trans. *Am. Soc. Civ. Eng.*, 116, 770-779.

47. Hurst H.E. (1951). Long-term storage of reservoirs: an experimental study. *Transactions of the American Society of Civil Engineers*, 116, 770-799.

48. Iqbal M. (1983). An introduction to solar radiation. Toronto. *Academic Press*, 1983.

49. Jacovides C.P., Tymvios F.S., Assimakopoulos V.D. and Kaltsounides N.A. (2006). Comparative study of various correlations in estimating hourly diffuse fraction of global solar radiation. *Renewable Energy*, 31, 2492 – 2504.

50. Javier E., Contreras R. and Wilfredo P. (2012). Statistical Analysis of Autoregressive Fractionally Integrated Moving Average Models. *Stat.*CO.

51. Jiang Y. (2008). Prediction of monthly mean daily diffuse solar radiation using artificial neural networks and comparison with other empirical models. *Energy Policy*, 36, 3833–3837.

52. Jiao C.L. (1990). System Theory of Artificial Neutral Network. *Publishing House of Xi'an University of Electron Technology, Xi'an*.

53. Kalekar P.S. (2004). Time series forecasting using Holt-Winters Exponential Smoothing. *Kanwal Rekhi School of Information Technology*.

54. Kansal A., Hsu J., Zahedi S. and Srivastava M.B. (2007). Power Management in Energy Harvesting Sensor Networks. ACM *Transactions on Embedded Computing Systems* 6.

55. Khatib T., Mohamed A., Sopian K. and Mahmoud M. (2012). Solar Energy Prediction for Malaysia Using Artificial Neural Networks. *International Journal of Photoenergy*.

56. Klein and Beckman (1987). *Energy Efficiency and Renewable Energy*. CRC Press.

57. Kobayashi Y., Watanabe R., Hida Y., Yokoyama R. and Funabashi T. (2013). Corrective Day-ahead Prediction of Solar Radiation based on Online Measurements. *Renewable Energies and Power Quality* (RE&PQJ).

58. Liu B.Y.H., and Jordan R.C. (1960). The Interrelationship and Characteristic Distribution of Direct, Diffuse and Total Solar Radiation. *Solar Energy* 4, 1-9.

59. Ljung G.M. and Box G.E.P. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65, 297–303.

60. Lucheroni C. (2007). Resonating models for the electric power market. *Physical Review* E 76, 9831.

61. Markvart T., Fragaki A. and Ross J.N. (2006). PV system sizing using observed time series of solar radiation. *Solar Energy*, 80, 46–50.

62. Newland F. J. (1988). A study of solar radiation models for the Coastal Region of South China. *Solar Energy* 31, 227–235.

63. Paoli C., Voyant C., Muselli M. and Nivet M. (2009). Solar radiation forecasting using ad-hoc time series preprocessing and neural networks. *Lecture Notes in Computer Science* 5754, 898-907.

64. Parton W.J. and Innis G.S. (1972). Some graphs and their function forms. US/IBP Grassland Biome Tech. Rep. No. 153. Colorado State Univ., Fort Collins. 41 pp.

65. Paulescu M., Paulescu E., Grauvila P. and Badescu V. (2012). Weather Modeling and Forecasting of PV Systems Operation. *Green energy and technology*, ISSN 1865-3529. Springer, 2012.

66. Perdomo R., Banguero E. and Gordillo G. (2010). Statistical modeling for global solar radiation forecasting in Bogota. *Photovoltaic Specialists Conference*, 35, 2374 –2379.

67. Pollock D.S.G., Richard C. and Nguyen T. (1999). A Handbook of Time Series Analysis, Signal Processing and Dynamics (Signal Processing and Applications). *Amazon.*

68. Prescott J.A. (1940). Evaporation from a water surface in relation to solar radiation. Trans. *Roy. Soc. South Australia* 64, 114–125.

69. Radosavljevic J. and Dordevic A. (2001). Defining of the intensity of solar radiation on horizontal and oblique surfaces on earth. *Working and Living Environmental Protection*, 2, 77 – 86.

70. Raja I.A. and Twidell J.W. (1990). Diurnal Variation of Global Insolation Over Five Locations in Pakistan. *Solar Energy*, 44, 73-76.

71. Reikard G. (2009). Predicting solar radiation at high resolutions: A comparison of time series forecasts. Solar Energy 83, 342-349.

72. Reisen V., Abrahem B. and Lopes S. (1993). Estimation of Parameters in ARFIMA Processes. *A simulation study*.

73. Samuel T.D. M.A. (1991). Estimation of global radiation for Srilanka. *Solar Energy*, 47, 333.

74. Schwarz G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

75. Srivastava R.C. and Pandey H. (2013). Estimating Angstrom-Prescott Coefficients for India and Developing a Correlation between Sunshine Hours and Global Solar Radiation for India. *ISRN Renewable Energy*.

76. Trouve A. Younes L. (2005). Local geometry of deformable templates. SIAM J. *Mathematical Analysis*.

77. Tukey J.W. (1957). On the comparative anatomy of transformations. *Annals of Mathematical Statistics*, 28, 602-632.

78. Tymvios F.S., Jacovides C.P., Michaelides S.C. and Scouteli C. (2005). Comparative study of Angstrom's and artificial neural networks methodologies in estimating global solar radiation. *Solar Energy*, 78, 752–762.

79. Ulgen K. & Hepbasli A. (2004). Solar Radiation Models. Part 2: Comparison and Developing New Models, Energy Sources, 26, 521-530.

80. Wang F., Mi Z., Su S. and Zhao H. (2012). Short-Term Solar Irradiance Forecasting Model Based on Artificial Neural Network Using Statistical Feature Parameters. *Energies*, 5, 1355-1370.

81. Watt A.D. (1978). "On the Nature and Distribution of Solar Radiation". *U.S. Department of Energy Report*, HCP/T2552-01.

82. Wold H. (1938). A Study in the Analysis of Stationary Time Series. *Almqvist and Wiksell, Sweden.*

83. Yeo I. and Johnson R.A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, 87, 954-959.

84. Zabara K. (1986). *Solar and Wind Technology*, 3, 267.

85. Zawilska E. and Brooks M.J. (2012). An Assessment of the Solar Resource for Durban, South. Africa. *Renewable Energy*, 36, 3433-3438.

86. Dazhi Y., Jirutitijaroen P. and Walsh W. M. (2012). Hourly solar irradiance time series forecasting using cloud cover index. *Solar Energy* 86, 3531–3543.

87. Martin L., Zarzalejo L. F., Polo J., Navarro A., Marchante R. and Cony M. (2010). Prediction of global solar irradiance based on time series analysis: Application to solar thermal power plants energy production planning. *Solar Energy* 84, 1772–178.