2016

Understanding Coupling of Global and Diffuse Solar Radiation with Climatic Variability

Lubna Hamdan

Follow this and additional works at: https://researchrepository.wvu.edu/etd

Recommended Citation
Hamdan, Lubna, "Understanding Coupling of Global and Diffuse Solar Radiation with Climatic Variability" (2016). Graduate Theses, Dissertations, and Problem Reports. 5742.
https://researchrepository.wvu.edu/etd/5742

This Dissertation is brought to you for free and open access by The Research Repository @ WVU. It has been accepted for inclusion in Graduate Theses, Dissertations, and Problem Reports by an authorized administrator of The Research Repository @ WVU. For more information, please contact ian.harmon@mail.wvu.edu.
UNDERSTANDING COUPLING OF GLOBAL AND DIFFUSE SOLAR RADIATION WITH CLIMATIC VARIABILITY

Lubna Hamdan

Dissertation submitted to the Statler College of Engineering and Mineral Resources at West Virginia University in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Industrial Engineering

Majid Jaridi, Ph.D., Chair
Antarpreet Jutla, Ph.D., Co-chair
E. James Harner, Ph.D.
Feng Yang, Ph.D.
Ashish Nimbarte, Ph.D.
Robert Mnatsakanov, Ph.D.

Department of Industrial and Management Systems Engineering

Morgantown, West Virginia
2016

Keywords: Global Radiation, Diffuse Radiation, Atmospheric parameters, Regression Analysis, Multivariate Analysis, Prediction Models

Copyright 2016 Lubna Hamdan
Abstract

**Understanding Coupling of Global and Diffuse Solar Radiation with Climatic Variability**

Lubna Hamdan

Global solar radiation data is very important for wide variety of applications and scientific studies. However, this data is not readily available because of the cost of measuring equipment and the tedious maintenance and calibration requirements. Wide variety of models have been introduced by researchers to estimate and/or predict the global solar radiations and its components (direct and diffuse radiation) using other readily obtainable atmospheric parameters. The goal of this research is to understand the coupling of global and diffuse solar radiation with climatic variability, by investigating the relationships between these radiations and atmospheric parameters. For this purpose, we applied multilinear regression analysis on the data of National Solar Radiation Database 1991-2010 Update.

The analysis showed that the main atmospheric parameters that affect the amount of global radiation received on earth’s surface are cloud cover and relative humidity. Global radiation correlates negatively with both variables. Linear models are excellent approximations for the relationship between atmospheric parameters and global radiation. A linear model with the predictors total cloud cover, relative humidity, and extraterrestrial radiation is able to explain around 98% of the variability in global radiation.

For diffuse radiation, the analysis showed that the main atmospheric parameters that affect the amount received on earth’s surface are cloud cover and aerosol optical depth. Diffuse radiation correlates positively with both variables. Linear models are very good approximations for the relationship between atmospheric parameters and diffuse radiation. A linear model with the predictors total cloud cover, aerosol optical depth, and extraterrestrial radiation is able to explain around 91% of the variability in diffuse radiation. Prediction analysis showed that the linear models we fitted were able to predict diffuse radiation with efficiency of test adjusted $R^2$ values equal to 0.93, using the data of total cloud cover, aerosol optical depth, relative humidity and extraterrestrial radiation. However, for prediction purposes, using nonlinear terms or nonlinear models might enhance the prediction of diffuse radiation.
DEDICATION

I dedicate this work to my husband Nader and to my family. Their support, encouragement, and prayers helped me overcome the difficulties and eased the pressure that I faced during my PhD study.
ACKNOWLEDGMENTS

I would like to express my sincere appreciation to my advisor, Dr. Majid Jaridi, for his guidance, continued help, support, and encouragement, and for having confidence in me. I would like to thank my co-advisor, Dr. Antarpreet Jutla, for his guidance, help and support throughout my research. I would like to thank Dr. E. James Harner for guiding me throughout the statistical analysis of my research. I was lucky to work with all of them and learn from their experience and skills in the field of research.

I would like to thank my committee members: Dr. Feng Yang, Dr. Ashish Nimbarte, and Dr. Robert Mnatsakanov for being my committee members and accepting the heavy task of going through my dissertation. I am grateful to all my professors, my colleagues, and the staff of Industrial Engineering Department for their help and support.

Finally, my thanks and appreciation are to my husband, my family, and my friends for their support, help, and encouragement. I am blessed to have such a supportive and loving family and friends.
Table of Contents

Abstract ................................................................................................................................. ii
DEDICATION ......................................................................................................................... iii
ACKNOWLEDGMENTS ........................................................................................................ iv
LIST OF FIGURES ................................................................................................................ vii
LIST OF TABLES .................................................................................................................. x
NOMENCLATURE ................................................................................................................ xi
CHAPTER 1: INTRODUCTION ............................................................................................... 1
CHAPTER 2: LITERATURE REVIEW ..................................................................................... 4
  2.1 Global Radiation ........................................................................................................... 4
  2.2 Diffuse Radiation ........................................................................................................ 18
CHAPTER 3: PROBLEM STATEMENT AND PROPOSED APPROACH ................................. 23
  3.1 Approach .................................................................................................................... 23
  3.2 Data ............................................................................................................................ 24
    3.2.1 Source of data ...................................................................................................... 24
    3.2.2 Data files ............................................................................................................. 27
    3.2.3 Data testing ......................................................................................................... 30
CHAPTER 4: GLOBAL RADIATION .................................................................................. 32
  4.1 Preliminary Analysis .................................................................................................... 32
  4.2 Multiple Linear Regression Analysis ......................................................................... 33
    4.2.1 Models with Clearness Index (K) response ......................................................... 34
    4.2.2 Models with log (H) response ........................................................................... 41
    4.2.3 Models with H response .................................................................................... 46
  4.3 Prediction Analysis ..................................................................................................... 50
  4.4 Discussion and Conclusions ....................................................................................... 53
CHAPTER 5: DIFFUSE RADIATION ................................................................................... 57
  5.1 Preliminary Analysis .................................................................................................... 57
    5.2.1 Models with Diffuse Fraction (Kd) Response ...................................................... 58
    5.2.2 Models with Diffuse Radiation (Hd) Response .................................................... 66
    5.2.3 Models with log (Hd) Response ......................................................................... 73
  5.3 Prediction Analysis ..................................................................................................... 76
5.4 Discussion and Conclusions ........................................................................................................ 78
CHAPTER 6: COMPLEMENTARY ANALYSES .................................................................................. 81
  6.1 Regularized Regression .............................................................................................................. 81
  6.2 Autoregressive Analysis for the Residuals ................................................................................. 83
  6.3 Leave Out One Station Cross Validation .................................................................................. 86
CHAPTER 7: GENERAL CONCLUSIONS AND FUTURE WORK .................................................... 88
  7.1 General Conclusions ................................................................................................................ 88
  7.2 Technical Contributions .......................................................................................................... 89
  7.3 Impact on Energy Industry ....................................................................................................... 89
  7.4 Future Work ............................................................................................................................. 89
APPENDIX ......................................................................................................................................... 91
REFERENCES .................................................................................................................................. 96
LIST OF FIGURES

Figure 1.1: Solar radiation components segregated by the atmosphere and surface-------------------2
Figure 3.1: NSRDB 1991-2010 stations with their classification--------------------------25
Figure 3.2: Daily Statistical File of NSRDB 1991-2010 for Albuquerque, New Mexico--------27
Figure 3.3: Pictorial description for cloud cover derivation-----------------------------29
Figure 3.4: Normal QQ plots for solar radiations and atmospheric parameters data--------30
Figure 3.5: Probability distributions for solar radiations and atmospheric parameters----31
Figure 4.1: The scatterplot matrix of global radiation and atmospheric parameters---------32
Figure 4.2: The Correlation matrix of global radiation and atmospheric parameters-------33
Figure 4.3: Best subset selection results for the models with K response----------------36
Figure 4.4: Best model of K response subsets based on adjusted R², Cp and BIC selection---36
Figure 4.5: The results of fitting K against OpqC, RH, H0, and TRange variables-----------37
Figure 4.6: Diagnostic plots for model (4.6). A- Linearity and constant variance of ε test.
   B- Normality test (ε~N(0, σ²)). C- Outliers test. D- High leverage points test-------38
Figure 4.7: Residuals vs Time. Left side is for FL station Data, Right side is for
   Corpus-TX station------------------------------------------------------------------------40
Figure 4.8: Best subset selection results for the models with log (H) response----------42
Figure 4.9: Best log (H) response model of best subsets based on adjusted R², Cp and
   BIC selection----------------------------------------------------------------------------42
Figure 4.10: The results of fitting log (H) against log (H₀), OpqC, RH, and TRange --------43
Figure 4.11: Diagnostic plots for model (4.11). A- Linearity and constant variance of ε test.
   B- Normality test (ε~N(0, σ²)). C- Outliers test. D- High leverage points test------44
Figure 4.12: The results of fitting log (H) against offset (log (H₀)), OpqC, RH, and TRange ----45
Figure 4.13: Best subset selection results for the models with H response----------------46
Figure 4.14: Best model of H response subsets based on adjusted R², Cp and BIC selection----47
Figure 4.15: The results of fitting H against H₀, TotC, RH, and T---------------------------------47
Figure 4.16: Diagnostic plots for model (4.13). A- Linearity and constant variance of $\epsilon$ test. 
B- Normality test ($\epsilon \sim N(0, \sigma^2)$). C- Outliers test. D- High leverage points test-----48

Figure 4.17: Residuals vs Time for H response model. Left side is for FL station Data, Right side is for Corpus-TX station-------------------------------------------------------------49

Figure 5.1: The scatterplots and correlations of average daily diffuse radiation and daily diffuse fraction with atmospheric parameters-----------------------------------------------57

Figure 5.2: Best subset selection results for the Kd response models----------------------------------58

Figure 5.3: Best model of Kd response subsets based on adjusted $R^2$, Cp and BIC selection-----59

Figure 5.4: Best subset selection results for the Kd response models without predictor K---------59

Figure 5.5: Best Kd response model of the subsets without K predictor, based on adjusted $R^2$, Cp and BIC selection---------------------------------------------------------------60

Figure 5.6: The results of fitting Kd against K, AOD, and RH variables---------------------------62

Figure 5.7: Diagnostic plots for model (5.1). A- Linearity and constant variance of $\epsilon$ test. 
B- Normality test ($\epsilon \sim N(0, \sigma^2)$). C- Outliers test. D- High leverage points test------63

Figure 5.8: Residuals vs Time for model (5.1). Left is for FL station Data, right is for 
Corpus-TX--------------------------------------------------------------------------------------63

Figure 5.9: The results of fitting Kd against OpqC, AOD, $H_0$, and RH variables------------------65

Figure 5.10: Diagnostic plots for model (5.3) -------------------------------------------------------65

Figure 5.11: Residuals vs Time for model (5.3) --------------------------------------------------------66

Figure 5.12: Best subset selection results for the models with Hd response-----------------------67

Figure 5.13: Best model of Hd response subsets based on adjusted $R^2$, Cp and BIC selection----67

Figure 5.14: The results of fitting Hd against $H_0$, H, AOD, and TotC-----------------------------68

Figure 5.15: Diagnostic plots for model (5.14). A- Linearity and constant variance of $\epsilon$ test. 
B- Normality test ($\epsilon \sim N(0, \sigma^2)$). C- Outliers test. D- High leverage points test------69

Figure 5.16: Residuals vs Time for model (5.4). Left side is FL station Data, Right side is 
Corpus-TX station----------------------------------------------------------------------------------70

Figure 5.17: Best subset selection results for Hd response models that exclude predictor $H_0$ ----70

Figure 5.18: The results of fitting Hd against H, AOD, TotC, and RH-------------------------------71
Figure 5.19: Best subset selection results for $H_d$ response models that exclude predictor $H$——72
Figure 5.20: The results of fitting $H_d$ against $H_0$, AOD, TotC, and RH-------------------72
Figure 5.21: Best subset selection results for $\log(H_d)$ response models-------------------74
Figure 5.22: The results of fitting $\log(H_d)$ against $\log(H_0)$, AOD, TotC, and RH--------74
Figure 5.23: Diagnostic plots for model (5.8) ------------------------------------------------75
Figure 5.24: Residuals vs Time for model (5.8). Left side is FL station Data, Right side is Corpus-TX station-----------------------------------------------75
Figure 6.1: Autoregressive function for residuals of model (4.12) --------------------------------85
Figure 6.2: Autoregressive function for residuals of model (4.13) --------------------------------85
LIST OF TABLES

Table 3.1: First class stations that have both measured diffusion and global radiation---------26
Table 3.2: Variables used in the analysis with their definitions and measure units---------------28
Table 4.1: The test data used to evaluate models performance---------------------------------51
Table 4.2: The accuracy of global models based on the statistical error tests------------------52
Table 4.3: The most important parameters in descending order for the major global models-----54
Table 5.1: The accuracy of diffuse models based on the statistical error tests-------------------76
Table 5.2: The most important parameters in descending order for the major diffuse models-----79
Table 6.1: The coefficients and performance of MLR models, the lasso ($\lambda=3.4$) and ridge regression ($\lambda=192.0$) for global radiation------------------------------------------84
Table 6.2: The coefficients and performance of MLR models, the lasso ($\lambda=0.06$) and ridge regression ($\lambda=57.86$) for diffuse radiation------------------------------------------84
Table 6.3: LOOSCV for model (6.3); station name represents the station used in testing the model---------------------------------------------------------------86
Table 6.2: LOOSCV for model (6.4); station name represents the station used in testing the model---------------------------------------------------------------87
NOMENCLATURE

AOD  Monthly average of broadband aerosol optical depth (unitless)
C    Cloud cover in oktas or tenths
C_p  Mallow’s C_p
\(d\)  Hourly diffuse fraction
E    Mean evaporation (cm)
F    Time of sunshine expressed in the greatest possible time of sunshine (sunshine fraction)
\(\bar{F}\)  Mean value of sunshine fraction
\(f_{\text{clear}}\)  Time fraction that no significant clouds block the sun
h    Elevation (m)
\(H\)  Monthly average of the daily global solar radiation on a horizontal surface (Wh/m^2)
\(H_b\)  Monthly average of the daily beam radiation on a horizontal surface (Wh/m^2)
\(H_d\)  Monthly average of the daily diffuse radiation on a horizontal surface (Wh/m^2)
\(H_0\)  Monthly average of daily extraterrestrial solar radiation on a horizontal surface (Wh/m^2)
\(H_{\text{clear}}\)  Monthly average of daily clear sky horizontal surface radiation (J/m^2)
\(H_2O\)  Monthly average of precipitable water (cm)
\(I_D\)  Diffuse horizontal irradiation (W/m^2)
\(I_G\)  Global horizontal irradiation (W/m^2)
\(I_{GC}\)  Clear-sky global horizontal irradiation (W/m^2)
\(I_N\)  Normal terrestrial beam solar irradiation (W/m^2)
\(I_{ON}\)  Normal extraterrestrial solar irradiation (W/m^2)
\(K\)  Monthly average of clearness index \(K = \frac{H}{H_0}\)
\(K_{\text{clear}}\)  Monthly average of clear sky clearness index \(K_{\text{clear}} = \frac{H_{\text{clear}}}{H_0}\)
\(K_d\)  Monthly average of diffuse fraction \(\left(\frac{H_d}{H}\right)\)
\(K_t\)  Daily clearness index
\(k_t\)  Hourly clearness index
### Variables and Definitions

- \( n \): Number of observations
- \( OpqC \): Monthly average of Opaque sky cover (tenths)
- \( p \): Number of predictors in a model
- \( Q \): Total radiation income during a day (MJ/m²)
- \( Q_{\text{clear}} \): Radiation income that corresponds to a perfectly clear sky day (MJ/m²)
- \( Q_0 \): Extraterrestrial radiation on a horizontal surface during the day (MJ/m²)
- \( R^2 \): Coefficient of determination
- \( RH \): Monthly average of relative humidity (%)
- \( S \): Monthly average daily bright sunshine hours
- \( S_0 \): Maximum possible monthly average daily sunshine hours or the day length
- \( ST \): Mean soil temperature (°C)
- \( T \): Monthly average of temperature or mean temperature (°C)
- \( T_{\text{Max}} \): Monthly average of maximum temperature (°C)
- \( T_{\text{Min}} \): Monthly average of minimum temperature (°C)
- \( T_{\text{Daylight}} \): Monthly average of daylight temperature (°C)
- \( T_{\text{Range}} \): Monthly average of temperature range (°C)
- \( TotC \): Monthly average of total sky cover (tenths)
- \( \Delta T \): Daily temperature range
- \( WS \): Monthly average of wind speed (m/s)
- \( x_{ij} \): \( i \)th observation for the variable \( j \)
- \( y_i \): \( i \)th measured response
- \( \hat{y}_i \): \( i \)th fitted or predicted response
- \( \bar{y} \): The average or mean of the responses \( Y \) (\( y_1 \) ... \( y_n \))
- \( \alpha \): Solar elevation (the angle between the horizon and the center of the sun's disc) (radian or (°)).
- \( \delta \): Solar declination (the angle between the rays of the Sun and the plane of the Earth's equator) (radian or (°)).
- \( \epsilon \): Error (the difference between measured and predicted responses)
\( \theta_z \)  Solar zenith angle (the angle between the vertical and center of the sun's disc) (radian or (°))

\( \lambda \)  Tuning parameter for ridge regression and the lasso

\( \hat{\sigma}^2 \)  Estimated variance of errors

AR  Autoregressive

BIC  Bayesian Information Criterion

ISIS  Integrated Surface Irradiance Study network

LOOSCV  Leave Out One Station Cross Validation

MAPE  Mean Absolute Percentage Error

MBE  Mean Bias Error

MLR  Multiple Linear Regression

MPE  Mean Percentage Error

MSE  Mean Squared Error

NCDC  National Climatic Data Center

NOAA  National Oceanic and Atmospheric Administration

NREL  National Renewable Energy Laboratory

NSRDB  National Solar Radiation Database

RMSE  Root Mean Squared Error

RSS  Residual Sum of Squares

SURFRAD  Surface Radiation Budget Measurement network

TSS  Total Sum of Squares

UO  University of Oregon Solar Radiation Monitoring Laboratory Network

USAF  United States Air Force

UT  University of Texas Solar Energy Laboratory

VIF  Variance Inflation Factor
CHAPTER 1: INTRODUCTION

Solar radiation arriving the earth’s surface is the most fundamental renewable energy source in nature. It sustains the biosphere and drives its self-organization; it also drives many of earth’s physical processes. Time and space dependent solar radiation changes the distribution of temperature, moisture, clouds, and precipitation as well as the pattern of atmospheric and oceanic circulations (Zhang et al, 2013). In addition, solar radiation, as a source of renewable energy, can play a key role in de-carbonizing the global economy since it is abundant and harnessing it has little adverse environmental impact. There is hardly any pollution in the form of exhaust fumes or even noise associated with conventional solar energy generation technologies. Accordingly, knowing the amount of solar radiation that reaches the surface of the earth is very important for a wide range of applications in engineering, meteorology, agricultural sciences, health sector, and natural sciences. Some examples of applications that use the solar radiation data at ground level include air conditioning and cooling systems in architecture and building design, solar heating system design and use, solar power generation and solar powered car races. As well, weather and climate prediction models, evaporation and irrigation, calculation of water requirements for crops, monitoring plant growth, and disease control and skin cancer research (Badescu, 2008).

Solar radiation reaching the earth’s surface is time and space dependent. A summary of the parameters affecting solar irradiation is as follows (Ertekin and Yaldiz, 1999):

- Astronomical factors (solar constant, earth-sun distance, solar declination and hour angle)
- Geographical factors (latitude, longitude and elevation of the site)
- Geometrical factors (azimuth angle of the surface, tilt angle of the surface, sun elevation angle, sun azimuth angle)
- Physical factors (water vapor content, scattering of dust and particulates, scattering of air molecules such as O₂, N₂, CO₂, O₃, etc.)
- Meteorological factors (effects of cloudiness, reflection of the environs)

We can calculate solar radiation incident on a horizontal plane outside the atmosphere (extraterrestrial radiation) at any point, using astronomical, geometrical, and geographical parameters of the site at a specific time. The details are in (Duffie and Beckman, 2013). However,
radiation incident on the earth’s surface, at some point, is random in nature due to the effect of physical and meteorological factors. Namely, as extraterrestrial radiation traverses the atmosphere, gases, dust, water vapor, and clouds within the atmosphere reflect, scatter and absorb the solar radiation at different wavelengths.

Due to the interaction between solar radiation and atmosphere constituents, we have two components of solar radiation incident on a horizontal plane at earth’s surface. The first component is diffuse sky radiation or simply diffuse radiation, which results from scattered photons (mostly at short wavelengths). The remaining unabsorbed and unscattered photons constitute the second component, direct beam radiation, which is responsible for the casting of shadows. The total solar radiation flux resulting from diffuse and beam radiations on a horizontal surface is called total or global solar radiation. The difference between global solar radiation at the top of the atmosphere and its corresponding value at the ground level is the amount of radiation that atmosphere has absorbed or reflected away. On average, earth reflects about 29% of the incident solar radiation. For a tilted surface, beside the beam and diffuse radiations, there is a third component, which is the radiation, reflected from the ground, see Figure 1.1 (Gueymard and Myers, 2008).

Figure 1.1: Solar radiation components segregated by the atmosphere and surface.

Pyrheliometer is the instrument that measures the direct beam radiation. Pyrheliometers have a narrow aperture (generally between 5° and 6° total solid angle), admitting only beam radiation with some inadvertent circumsolar contribution from the Sun’s aureole. However, the
aperture excludes all diffuse radiation from the sky. Pyrheliometers must be pointed at sun and
track it through the day. Their sensor should be always normal to the direct beam. Pyranometer is
the instrument that measures the global radiation or the diffuse sky radiation. Pyranometers have
a 180° (2π steradian) field of view. To measure diffuse radiation using the pyranometer, beam
radiation is blocked out with a disk or ball placed over the instrument and in the path of the direct
beam. The blocking device must track the sun through the day (Gueymard and Myers, 2008).

Unfortunately, in spite of the importance of solar radiation measurements, these data are
not readily available for many developing countries because of the cost of measuring equipment
and the tedious maintenance and calibration requirements (Al-Mostafa et al, 2014). Even all over
the world, weather stations measuring solar radiation are very sparse. For example, in USA, 1%
of meteorological stations are recording solar radiation. In China, only 122 stations are recording
solar radiation out of more than 2000 stations have records of meteorological data. In Australia,
the ratio of weather stations recording global solar radiation to those recording air temperatures
was approximately 16 to 845 in 2006. Worldwide, the ratio of stations recording solar radiation to
those recording temperature is about 1:500 (Chen and Li, 2012). Because of the limited coverage
of solar radiation measuring networks, there is a need for developing solar radiation models able
to estimate the data required for solar-energy applications. Since 1920’s a number of methods and
correlations have been developed to estimate global solar radiation, based on the more readily
available meteorological data (sunshine duration, cloud cover, temperature...etc.) at the majority
of weather stations. However, these models depend on the local geographical, physical, and
meteorological factors of the site of interest. Next, we present a brief review for some of these
models.
CHAPTER 2: LITERATURE REVIEW

2.1 Global Radiation

Angstrom (1924) developed the first basic model for estimating global solar radiation when he introduced his famous empirical equation, which relates the global solar radiation on a horizontal surface, scaled by clear-sky global radiation, to the sunshine fraction:

\[
\frac{Q}{Q_{\text{clear}}} = a + (1 - a)F \tag{2.1}
\]

where \(Q\) is the total radiation income during a day (MJ/m\(^2\)/day), \(Q_{\text{clear}}\) is the radiation income that corresponds to a perfectly clear sky day (MJ/m\(^2\)/day), \(F\) is the time of sunshine expressed in the greatest possible time of sunshine, and \(a\) is empirical coefficient. Angstrom obtained the value \(a = 0.25\) for Stockholm.

Prescott (1940) suggested using a modified form of Angstrom equation (Eq. (2.1)), since it is difficult to estimate \(Q_{\text{clear}}\). The modification is to use the radiation incident on a horizontal surface with transparent atmosphere (extraterrestrial radiation) instead of cloudless day:

\[
\frac{Q}{Q_0} = a + bF \tag{2.2}
\]

where \(Q_0\) is the extraterrestrial radiation on a horizontal surface during the day (MJ/m\(^2\)/day), “\(a\)” and “\(b\)” are coefficients that depend on the location.

Equation (2.2) is known as Angstrom-Prescott equation or model.

Hereafter, hundreds of articles in the literature introduced new models and improvements on the existing models, including the techniques used, to improve solar radiation estimation using readily available meteorological variables. Next, is a selection of these models presented in chronological order.

Black et al. (1954) used data collected from 32 stations around the world to estimate the coefficients of Angstrom-Prescott equation (Eq. (2.2)). They obtained the following general equation (at least within the range of latitudes studied) for predicting solar radiation from sunshine duration:
\[
\frac{H}{H_0} = 0.23 + 0.48 \frac{S}{S_0}
\]  

(2.3)

where \(H\) is the monthly average of the daily global solar radiation on a horizontal surface (MJ/m\(^2\)/day), \(H_0\) is monthly average of daily total insolation on an extraterrestrial horizontal surface. \(S\) is the monthly average daily bright sunshine hours and \(S_0\) is the maximum possible monthly average daily sunshine hours or the day length \((S/S_0 = F\) in Equations (2.1) and (2.2)).

However, they pointed out that there are errors in the data they did not consider: the different periods of collection for the stations and the different instruments used in different countries.

Rietveld (1978) examined several published values of Angstrom-Prescott coefficients and found that \((a)\) is related linearly with \(\bar{F}\) (the mean value of fraction of sunshine) and \((b)\) is related hyperbolically with \(\bar{F}\) \((b \propto 1/\bar{F})\). His analysis showed that the use of these relationships to establish values of \(a\) and \(b\) provides more accurate estimates of radiation, from sunshine data, than does Black et al. (1954) method or any extrapolated use of existing formulae.

Kasten and Czeplak (1980) investigated the dependence of total solar and terrestrial radiation fluxes at the earth surface on cloud amount and cloud type. They used 10 years of hourly data of solar and terrestrial radiations and of cloud amount and type. In their analysis, they used solar elevation (the angle between the horizon and the center of the sun's disc) as a parameter instead of hour of the day.

Wahab (1993) derived a quadratic form of Angstrom-Prescott equation, based on a simple model relating global solar radiation to cloud amount and transmissivity, ground albedo, and atmospheric backscatter (Davies and McKay, 1989). Abdel Wahab obtained the following equation, with “c” coefficient always negative:

\[
\frac{H}{H_0} = a + b \frac{S}{S_0} + c \left(\frac{S}{S_0}\right)^2
\]  

(2.4)

Gueymard et al. (1995) criticized Wahab (1993) paper for the confusion between Angstrom equation (Eq. (2.1)) and Angstrom-Prescott equation (Eq. (2.2)), where Abdel Wahab analysis would be valid if he used \(Q_{\text{clear}}\) instead of \(Q_0\), since he made the derivation based on Angstrom equation. Furthermore, they pointed out a number of errors in the paper, which preclude its use in actual solar radiation calculations. In addition, they criticized the interpretation of other
researchers for Rietveld model (1978), where they used \( F \) (the monthly average of daily sunshine fraction) instead of \( \bar{F} \) (the yearly average of daily sunshine fraction) to estimate the coefficients in Angstrom-Prescott equation. Moreover, Gueymard et al. emphasized that Angstrom-Prescott model has received considerable attention, yet it is still highly empirical. They advised to concentrate on improving original Angstrom model (Eq. (2.1)) by explicitly relating its coefficients to climatological variables.

Ododo et al. (1996) correlated global solar radiation with cloud cover and sunshine duration fraction using the data of three Nigerian stations. They used the following relation to predict global solar radiation:

\[
\frac{H}{H_0} = b_0 + b_1 \frac{S}{S_0} + b_2 C + b_3 C \frac{S}{S_0} \quad (2.5)
\]

where \( C \) is the cloud cover in oktas. They obtained an excellent fit for one station and satisfactory results for the others.

Sen (1998) used theory of fuzzy sets to model solar irradiation and sunshine duration. He used a fuzzy logic algorithm for estimating the solar irradiation from sunshine duration measurements. The fuzzy model used does not provide an equation but can adjust itself to any type of linear or nonlinear form through fuzzy subsets of linguistic solar irradiation and sunshine duration variables. Sen believed that this method is suitable because solar radiation is a random process. He applied this method on some stations in the western part of Turkey. However, fuzzy logic algorithm may give better estimations, but it does not give physical interpretations for the results as regression analysis does.

Ertekin and Yaldiz (1999) used multiple linear regression models to estimate the monthly average daily global radiation for Antalya, Turkey using nine different variables. The variables are extraterrestrial radiation, solar declination, ratio of sunshine duration, mean relative humidity, mean temperature, mean soil temperature, mean cloudiness, mean precipitation and mean evaporation. From these variables they constructed 511 equations and found that the best model is the one which contains the nine variables at \( r = 0.99861 \).
\[ H = -13.08 + 0.386H_0 + 0.902\delta + 0.2254RH + 11.59 \frac{S}{S_0} - 0.034T - 0.251ST - 0.977C - 0.0072H_2O + 0.2373E \quad (2.6) \]

where \( \delta \) is solar declination (the angle between the rays of the Sun and the plane of the Earth's equator) (\(^\circ\)), \( RH \) is mean relative humidity (%), \( T \) is mean temperature (\(^\circ\)C), \( ST \) is mean soil temperature (\(^\circ\)C), \( C \) is mean cloudiness (1-10), \( H_2O \) is mean precipitation (cm), and \( E \) is mean evaporation (cm).

However, they had excellent values of correlation coefficient (r), at least for one model of each kind. The values started from \( r = 0.98447 \) using one variable equation, to \( r = 0.99860 \) using eight variables. Adding more variables did not give substantial improvement in radiation predictions since some variables are dependent on each other.

Suehrcke (2000) mentioned that the cloud transmittance depends on the radiation average. Hence, it is not constant as assumed by Angstrom’s equation, which suggests a non-linear sunshine-radiation relationship. Suehrcke used the properties of instantaneous solar radiation to derive a non-linear sunshine-radiation relationship free from empirical parameters, namely:

\[ f_{\text{clear}} = \left( \frac{K}{K_{\text{clear}}} \right)^2 = \left( \frac{H}{H_{\text{clear}}} \right)^2 \quad (2.7) \]

where \( f_{\text{clear}} \) is time fraction that no significant clouds block the sun, \( H_{\text{clear}} \) is monthly average of daily clear sky horizontal surface radiation (\( J/m^2 \)), \( K \) is monthly average daily clearness index \((K = \frac{H}{H_0})\) and \( K_{\text{clear}} \) is monthly average clear sky clearness index \((K_{\text{clear}} = \frac{H_{\text{clear}}}{H_0})\).

Suehrcke believes that Equation (2.7) may be universally valid and Angstrom–Prescott equation is a local (linear) approximation of his equation.

Ertekin and Yaldiz (2000) validated 26 models, available to predict the monthly average daily global radiation on a horizontal surface, using an independent data set for Antalaya Turkey. The models include linear, quadratic, third order polynomial and exponential equations. In addition, the models have different climatological and geographic parameters. Their analysis showed that the third order polynomial of Angstrom-Prescott type is the most accurate model:
\[
\frac{H}{H_0} = a + b \left( \frac{S}{S_0} \right) + c \left( \frac{S}{S_0} \right)^2 + d \left( \frac{S}{S_0} \right)^3 \tag{2.8}
\]

Muneer and Gul (2000) evaluated the performance of Page radiation model, which combines clouds and sunshine data to predict solar radiation, against two models developed by the authors. The first model is Meteorological Radiation Model, which uses hourly dry and wet bulb temperatures, and sunshine fraction to estimate hourly global, beam and diffuse irradiation. The second is Cloud Cover Radiation Model, which uses the hourly data of cloud amount to predict solar radiation. For the evaluation, they used data from UK sites. The analysis showed that Page model performs with maximum efficiency under overcast conditions and the Meteorological Radiation model gives the best results under clear skies. For intermediate conditions, both the Meteorological and Cloud Cover models are capable of generating quality data.

Driesse and Thevenard (2002) tested Suehrcke’s equation (Eq. (2.7)) for the calculation of monthly average daily radiation on a horizontal surface. They used 70,000 measured monthly sunshine and radiation data from nearly 700 sites compiled by the World Radiation Data Center. They also compared the performance of Suehrcke’s model with Angstrom-Prescott Model (Eq. (2.2)). They concluded that Suehrcke’s equation accounts adequately for the sunshine–radiation relationship on an average sense. However, the predictive capabilities of Suehrcke model are actually roughly equivalent to those of Angstrom-Prescott model when the peculiarities of local climatic conditions are not considered.

Yorukoglu and Celik (2006) conducted a literature survey and showed that researchers investigate either the goodness of the Angstrom–Prescott equation type model itself or the goodness of the estimation of global solar radiation. If the former is the objective, then the statistical analysis should be based on the variables \(H/H_0\) and \(S/S_0\). If the investigation was for goodness of estimation, then the statistical analysis should be based on \(H_c\) and \(H_m\) (\textit{calculated daily solar radiation vs. measured daily solar radiation}). They showed that these two data sets are apt to be confused, where some researcher use \(H_c\) and \(H_m\) to investigate the goodness of the model or vice versa set. In addition, the authors compared five different Angstrom-Prescott type models (linear, quadratic, cubic, logarithmic and exponential) based on six years of measured hourly global solar radiation data. The analysis showed that amongst the five different models the Angstrom–Prescott equation, the quadratic and the cubic models are the best. Even though the
cubic model has slightly better performance than Angstrom–Prescott model (the simplest one), the advantage of the cubic model may be abandoned in return for a simpler model with half of the parameters.

Ertekin and Evrendilek (2007) compared the performance of eighteen empirical models in linear, quadratic, cubic, logarithmic, exponential and hybrid forms using only sunshine hours, latitude, and altitude. The models to estimate monthly average daily global solar radiation on a horizontal surface for 159 weather stations in Turkey. They found that the best models are a linear model (Angstrom type equation) with the coefficients that depend on the altitude and S/So and a hybrid model (quadratic with coefficients that depend on latitude and altitude). In addition, they generated spatial variability maps of global solar radiation on a 500 m x 500 m grid using the data of the 159 weather stations. However, they have values for $R^2_{adj}$ greater than $R^2$.

Mellit et al. (2007) developed a new approach for predicting and modelling of daily total solar radiation data from sunshine duration and air temperature. They used an Adaptive Neuro-Fuzzy Inference Scheme (ANFIS) model. They built the simulation model in Matlab, using ten-year database of daily sunshine duration, ambient temperature and total solar radiation data. They validated the model with unknown data and showed that its estimations were excellent. Compared with other Adaptive Neuro Network models, their model was the best and the faster. This paper used simulation technique instead of regression analysis.

Younes and Muneer (2007) compared seven solar radiation models based on cloud information. These models are M1: Kasten and Czeplak (1980) model, represented in the following equations:

\[ I_{GC} = 910 \sin \alpha - 30 \]  \hspace{1cm} (2.9)

\[ \frac{I_G}{I_{GC}} = 1 - 0.75 \left( \frac{C}{8} \right)^{3.4} \]  \hspace{1cm} (2.10)

\[ \frac{I_D}{I_G} = 0.3 + 0.7 \left( \frac{C}{8} \right)^2 \]  \hspace{1cm} (2.11)

where $I_{GC}$ is clear-sky global horizontal irradiation (W/m$^2$), $\alpha$ is solar elevation (radian), $I_G$ is global horizontal irradiation (W/m$^2$), $C$ is cloud cover (oktas), and $I_D$ is diffuse horizontal irradiation (W/m$^2$). M2: Local coefficient modified Kasten and Czeplak (Muneer and Gul, 2000).
(Gul et al., 1998), where the authors modified the coefficients of Equations (2.9) and (2.10) to fit the local data. **M3:** Lam and Li (1998) model, represented in the following equations:

\[
I_G = 217 - 485 \left( \frac{C}{8} \right) + 696\sin \alpha \quad (2.12)
\]

\[
I_D = 30.5 - 62.9 \left( \frac{C}{8} \right) + 294.7\sin \alpha \quad (2.13)
\]

**M4:** Local coefficient modified Lam and Li, where the authors modified the coefficients of Equations (2.12) and (2.13) based on the local data. **M5, M6** and **M7** new models proposed by the authors and represented by the following equations:

**M5:**

\[
I_G = I_{GC}(a_0 + a_1\varphi + a_2\varphi^2)^{b_0} \quad (2.14)
\]

\[
I_D = I_G(c_0 + c_1\varphi + c_2\varphi^2)^{d_0} \quad (2.15)
\]

**M6:**

\[
I_G = I_{GC}(a_0 + a_1\varphi + a_2\varphi^2)^{(b_0+b_1\varphi)} \quad (2.16)
\]

\[
I_D = I_G(c_0 + c_1\varphi + c_2\varphi^2)^{(d_0+d_1\varphi)} \quad (2.17)
\]

**M7:**

\[
I_G = I_{GC}(a_0 + a_1\varphi + a_2\varphi^2)^{(b_0+b_1\varphi+b_2\varphi^2)} \quad (2.18)
\]

\[
I_G = I_{GC}(c_0 + c_1\varphi + c_2\varphi^2)^{(d_0+d_1\varphi+d_2\varphi^2)} \quad (2.19)
\]

where \( \varphi = \frac{C}{8} \)

The analysis showed that the **M2** and **M7** models performed the best. For diffuse radiation, **M7** performed slightly better than **M2** Model.

Akinoglu (2008) made analytical review for the models that predict solar radiation based on sunshine duration in the literature. He explained the physical meaning of Angstrom-Prescott equation. Akinoglu discussed two broadband spectral physical models: The Hybrid model and the direct approach to physical modeling. The two models have different approaches but both reached
to the same results, that is, a quadratic relationship between fractional solar radiation and fractional sunshine duration.

Bakirci (2009) reviewed sixty global solar radiation models based on sunshine duration data. These models consist of relations derived from the Angstrom-Prescott equation. Bakirci categorized the models into four groups: 1- Linear models (first order regression analysis). 2- Polynomial models (second, third and larger order polynomial equations). 3- Angular models (contain trigonometric functions). 4- Other models (including logarithmic term, exponential term and non-linear terms). He concluded that the models presented in his study might be used reasonably well for estimating the solar radiation at a given location and possibly in elsewhere with similar climatic conditions.

Benghanen et al. (2009) developed artificial neural network (ANN) models for estimating daily global solar radiation. They used four years’ data of global irradiation, sunshine duration, air temperature, relative humidity, and the day of the year. They constructed six ANN models using different combinations as input with daily global solar radiation as the output. The analysis showed that the best is the model with the inputs of sunshine duration and air temperature. In addition, sunshine duration plays a very important role for obtaining high accurate results; where all models that have sunshine duration in their input have correlation coefficient greater than 97%. They also compared the models with the classical regression methods and again the best was the ANN model with sun duration and temperature as inputs. However, the classical model with quadratic correlation between H/Ho and S/So has very close result to that of the best ANN model.

Reikard (2009) evaluated the ability of several types of time series models to predict radiation at ground level using six data sets. Three consist of hourly time series from the National Solar Radiation Database for the locations Kansas City, MO, Denver, CO, and Phoenix, AZ. These series run from January 1, 1987 to December 31, 1990. The others are from the Measurements and Instrumentation Data Center and are at a basic resolution of 1 min. Reikard averaged the basic 1-min data to create time series at resolutions of 5, 15, 30, and 60 min. The evaluated models are regressions in logs, Autoregressive Integrated Moving Average (ARIMA), Unobserved Components Models (model diurnal cycle trigonometrically), Transfer functions (add causal inputs such as cloud cover), neural networks, and hybrid models (combined regressions and neural nets). The best results were for the ARIMA in logs, with time-varying coefficients. At high
resolutions, a transfer function using cloud cover improved over the ARIMA. In a few cases, the neural net or hybrid models could improve at very high resolutions, in the order of 5 min.

Lee (2010) modified the Angstrom-Prescott equation to a non-linear relationship between the incoming shortwave solar radiation and bright sunshine duration:

\[ \frac{Q}{Q_0} = a + b F^c \quad (2.20) \]

He used the data of solar radiation and sunshine radiation from 1997 to 2006 at 21 meteorological stations in Korea to calibrate and validate the suggested equation. He obtained a value of \( c = 0.649 \), that is \( c < 1 \). A comparison between the results of his model and Angstrom-Prescott model showed that the modified model performance is better. However, there is no significant difference between the two models.

Ahmad and Tiwari (2011) reviewed solar radiation models for predicting the average daily and hourly global radiation, beam radiation and diffuse radiation on horizontal surface. They divided the models to Parametric Models that require detailed information on atmospheric conditions (such as clouds, fractional sunshine and atmospheric turbidity) and Decomposition Models that usually use information only on global radiation to predict the beam and sky components. They discussed the following categories of models:

- Parametric models estimating hourly global irradiation. For the composite climate of India, the best model is Ahmad and Tiwari (2008) model:

\[
\begin{align*}
I_N &= I_{ON} \times \exp[(m\varepsilon)^2 T_{RO} + m\varepsilon T_R + \tau] \\
I_D &= K_0((I_{ON} - I_N) \cdot \cos \theta_z)^2 + K_1(I_{ON} - I_N) \cdot \cos \theta_z + K_2
\end{align*}
\]

where \( I_N \) is normal terrestrial beam solar irradiation (Wm\(^{-2}\)), \( I_{ON} \) is normal extraterrestrial solar irradiation, \( m \) air mass (dimensionless), \( \varepsilon \) is integrated Rayleigh scattering optical thickness, \( T_{RO} \) and \( T_R \) are cloudiness/haziness factor, \( \tau \) is atmospheric transmittance for beam radiation, \( I_D \) is diffuse solar irradiation (Wm\(^{-2}\)), and \( \theta_z \) is solar zenith angle (the angle between the vertical and center of the sun’s disc). The authors interpreted \( K_0, K_1 \) and \( K_2 \) as atmospheric transmittances for diffuse radiation.

- Decomposition models estimating hourly diffuse radiation on horizontal surface. They presented 14 models of this category without evaluating their performance.
• Models predicting the mean hourly global radiation from daily summations. The best model is Collares-Pereirs and Rabl model as modified by Gueymard (1986) (CPRG):

\[ r_{CPRG} = (a + b \cos \omega) r_0 / f \]

where \( r_0 \) is extraterrestrial radiation hourly/daily ratio, \( r_{CPRG} \) is a modification of \( r_0 \) to account for the atmospheric effect and ensure consistency through normalization and \( \omega \) is hour angle (an angular measure of time, equivalent to 15°/h, relative to solar noon, where solar noon hour angle = 0.00°). \( r_0 \), \( a \), \( b \) and \( f \) are functions of \( \omega_0 \) (sunrise angle hour), given in the following equations:

\[ r_0 = (\cos \omega - \cos \omega_0) / k A(\omega_0) \]

where \( A(\omega_0) = \sin \omega_0 - \omega_0 \cos \omega_0 \)

\[ a = 0.4090 + 0.5016 \sin \left( \omega_0 - \frac{\pi}{3} \right) \]

\[ b = 0.6609 - 0.4767 \sin \left( \omega_0 - \frac{\pi}{3} \right) \]

\[ f = a + 0.5b(\omega_0 - \sin \omega_0 \cos \omega_0) / A(\omega_0) \]

• Models correlating average daily global radiation with hours of sunshine. They presented fifty models of this category. Menges et al. (2006) evaluated the performance of these fifty models against data of Konya, Turkey. They found that Ertekin and Yaldiz (1999) model (Equation (2.6)) has the best performance.

Li et al. (2011) studied the significance of seven different solar constant values collected from literature for estimating the monthly average daily global solar radiation with Angstrom-Prescott correlation. The authors used measured data between 1971 and 2000 at eight meteorological stations in China, covering a diverse range in climate and geography. They fitted the coefficients of Angstrom-Prescott equation using the seven different values of solar constant. They evaluated the effect of the solar constant values on the Angstrom-Prescott correlation using a ranking method based on the t-statistic. The authors found that the results of all of them have significant meaning, but different places have different best value of solar constant. They proposed using different solar constants for different regions based on their climate.

Matuszko (2012) analyzed the influence of cloudiness and cloud genera on sunshine duration using very long (1884–2007) daily nephological and sunshine duration data for the City of Krakow (Poland). He used quadratic regression to describe the relationship between sunshine
duration in hours and cloud cover percentage. Analysis showed that cloudiness affects sunshine duration the most in June and July, and the least in December, January and February. High clouds (Cirrus, Cirrostratus, and Cirrocumulus) do not interrupt the recording of sunshine duration even when they completely cover the sky. Layered clouds such as Stratus and Nimbostratus do not transmit solar radiation at all. The influence of different cloud genera on sunshine duration changes minimally from season to season and with respect to the position of the Sun over the horizon. When the Sun position is high in the sky, clouds are less able to weaken solar radiation, resulting in larger sunshine duration values. This is especially true with respect to Cirrus, Cirrostratus and Cumulus clouds.

Chen and Li (2012) conducted a simple procedure to map the daily solar radiation for Liaoning province in China, which has sparse data of solar radiation. They interpolated the daily sunshine duration to the whole area and then calculated daily solar radiation by Angstrom-Prescott model. They fitted the model parameters using local available data (three stations have both solar radiation and sun duration). The interpolation of sunshine duration data was by using ANUSPLIN software. However, substitution of solar radiation from nearby station is preferred if the distance between the stations falls below the threshold of 135 ± 15 km.

Suehrcke et al. (2013) re-examined the relationship between sunshine duration and solar radiation received on the Earth’s surface, using the same data used by Driesse and Thevenard (2002). They developed a procedure to reject physically questionable data and analyzed sunshine-radiation data for a wide range of climates. Based on their analysis, they proposed a generalized nonlinear parameter free model, where Suehrcke equation (Eq. (2.7)) and Angstrom-Prescott equation are special cases of this model:

\[
\frac{K}{K_{\text{clear}}} = \beta + (1 - \beta) \left( \frac{S}{S_0} \right)^\gamma
\]

(2.28)

where \( K \) is monthly average of daily clearness index \( (K = \frac{H}{H_0}) \), \( K_{\text{clear}} \) is monthly average of daily clearness index for a cloudless day \( (K_{\text{clear}} = \frac{H_{\text{clear}}}{H_0}) \), and \( \beta, \gamma \) are constants.

The authors introduced other evidences from the literature for the nonlinearity between sunshine duration and radiation. They explained the cause of nonlinearity by the dependence of
clouds transmittance on sunshine fraction, where clouds become optically thicker (less transparent) with decreasing sunshine fraction.

Katiyar and Pandey (2013) presented a review of 61 global solar radiation models, from 1960 to 2010. The review considered Angstrom-Prescott type (linear) models, models of high order correlation, multi linear regression models, and models estimating global solar radiation based on ambient temperature. Their analysis showed that the second and third order correlations do not significantly improve the accuracy of the estimated global solar radiation over first-order. Accordingly, Angstrom-Prescott type correlation supersedes the second and third order correlations because of its accuracy and the less computational work it requires.

Zhang et al. (2013) developed an improved parametric model to estimate direct surface shortwave radiation on tilted surfaces under cloudy sky conditions. The improved parametric model integrates atmospheric attenuating effects with the three-dimensional effects correction of clouds and the topographic influences. The model estimates direct solar radiation in complex terrain under all sky conditions. To validate the model, they used MODIS (MODerate Resolution Imaging Spectroradiometer) satellite data of regions with different atmospheric conditions and surface roughness (Lhasa, Beijing, Kunming and Erjinaqi) in China. The data includes sensor zenith/azimuth, total water vapor, cloud optical thickness, cloud fraction and total atmospheric ozone. The results showed that the new parameterized model is convincingly efficient, as the computed coefficients of determination ($R^2$) are relatively high for all stations (the average around 0.7). Consequently, the model is a good estimator of the solar radiant energy for all sky and roughness conditions. In addition, since the input data are solely from the satellite products, the model is versatile and is not climate dependent.

Besharat et al. (2013) collected and reviewed the extensive global solar radiation models available in the literature and classified them into four categories: sunshine based models, cloud based models, temperature based models and other meteorological parameters based models. Then they selected several models from each category and evaluated their accuracy and applicability for computing the monthly average daily global solar radiation on a horizontal surface, using the geographical and meteorological data of the city Yazd in Iran. They compared the developed (calibrated) models based on statistical error indices and chose the most accurate model in each
category. They found that all the proposed models have a good estimation of the global solar radiation in Yazed with the El Metwally sunshine based model having the highest accuracy:

$$\frac{H}{H_0} = a \left( \frac{S}{S_0} \right)$$

(2.29)

The authors emphasized that global solar radiation models based on air temperature could be an important alternative to sunshine based models, in the absence of sunshine duration data, especially for locations with large temperature range.

Kacem Gairaa and Yahia Bakelli (2013) made a comparison between seven models for estimating the global solar radiation from sunshine duration, air temperature and relative humidity. The first four models are sunshine duration based models with linear, second order polynomial, logarithmic and exponential equations respectively. The following equations represent the last three models:

$$\frac{H}{H_0} = a + b \frac{S}{S_0} + c T_{max} + d (RH)$$

(2.30)

$$\frac{H}{H_0} = a + b \frac{S}{S_0} + c \left( \frac{T_{min}}{T_{max}} \right) + d \left( \frac{RH_{min}}{RH_{max}} \right)$$

(2.31)

$$\frac{H}{H_0} = a + b (T_{max} - T_{min})^{0.5}$$

(2.32)

They validated the models using a dataset of Ghardaia area in the south of Algeria. The analysis showed that the linear and quadratic models are the most suitable for estimating the global solar radiation from sunshine duration. For models based on meteorological parameters, Equations (2.30) and (2.31) give the best performance.

Yao et al. (2014) compared and analyzed 89 existing monthly average daily global solar radiation models and 19 existing daily global solar radiation models using 42-year (1961 to 2002) meteorological data. The results showed that for the existing monthly average daily global solar radiation models, linear and polynomial models were able to estimate global solar radiation accurately, while complex equation types could not improve the precision. Considering direct parameters such as latitude, altitude, solar elevation and sunshine duration can help improve the accuracy of the models, but indirect parameters such as relative humidity and maximum
temperature cannot. For existing daily global solar radiation models, multi-parameter models are more accurate than single parameter models and polynomial models are more accurate than linear models. In addition, the authors used the 42-year meteorological data to fit monthly average daily global solar radiation models based on sunshine duration. These models are linear, polynomial, logarithmic, exponential and power. They used the same data to fit daily global solar radiation models. The fitted models are linear, polynomial, power and exponential. Finally, the authors used 10 years (2003 to 2012) meteorological data, to compare existing models and fitting models. The results showed that polynomial models are the most accurate models.

Guclu et al. (2014) proposed a new model called dependency model. The basic idea of this model comes from Angstrom-Prescott approach with temporal dynamical effects between successive measurements:

\[
\left(\frac{H}{H_0}\right)_t = a + b \left(\frac{S}{S_0}\right)_t - d \left(\frac{S}{S_0}\right)_{t-i} + c \left(\frac{H}{H_0}\right)_{t-i}
\]

(2.33)

where \(i\) indicates the lag between the two time instants considered, \(c\) is the dependency coefficient for solar radiation and \(d\) is the dependency coefficient for sunshine.

The authors compared their model (dependency) with Angstrom-Prescott model (linear model) and the Adaptive Neuro Fuzzy Inference System that used to train input and output parameters. They used data of three southern cities in Turkey from 2000 to 2008. The analysis showed that the dependency model is superior over other approaches.

Lee (2014) developed Angstrom-Prescott equation and the modified Angstrom-Prescott equation (Eq. (2.20)) by adding the daily temperature range (DTR) to them. He added this term in an attempt to include the advection effect of meteorological variables (RH, T, cloudiness, wind speed, vapor pressure).

\[
\frac{Q}{Q_0} = a + b F^c + d(\Delta T)^e
\]

(2.34)

where \(c < 1\) in modified Angstrom-Prescott equation and \(\Delta T\) is daily temperature range (DTR).

Lee used the daily data of twenty meteorological stations in Korean Peninsula (1997 to 2006) to compare the four models: Angstrom-Prescott model, Angstrom-Prescott model + DTR,
modified Angstrom-Prescott model, and modified Angstrom-Prescott model + DTR. The results showed that adding DRT enhanced both models and the best is the modified Angstrom-Prescott model + DTR.

Moradi et al. (2014) evaluated six models that estimate the solar radiation. One model uses sunshine duration as its input (the Angstrom–Prescott model) and the other five models use the daily temperature range as their main input. They compared the models’ performance using data measured at four independent worldwide networks. The dataset included 13 stations from Australia, 25 stations from Germany, 12 stations from Saudi Arabia, and 48 stations from the USA. The results showed that Angstrom-Prescott model and the model of Bristow and Campbell (1984), see Equation (2.35), indicated a better performance than the other models.

\[
\frac{Q}{Q_0} = a \left(1 - e^{-b \Delta T(j)} \right) ^c \tag{2.35}
\]

where \( j \) is the day number from 1 to 366.

Polo et al. (2015) developed a model inspired by Angstrom equation, that is, it uses clear sky global solar radiation and sunshine duration. They estimated daily global horizontal radiation under clear sky conditions for 171 ground stations for the period 2003–2012 by using REST2 (Reference Evaluation of Solar Transmittance, 2-bands). They used canonical correlation analysis to fit the measured and created data of 11 radiometric stations for the period 2003-2012. The variables involved in the model are the daily global radiation, the daily global radiation under clear sky conditions, and the product of the daily extraterrestrial global radiation with the relative sunshine duration. The resultant model consists of four cubic polynomials, corresponding to each trimester of the year. They used the output of the model for characterizing the dispersion of long-term solar radiation and built spatial distribution for the variability of long-term solar radiation by clustering technique.

2.2 Diffuse Radiation

Since diffuse radiation is the component of global radiation, which is most affected by atmospheric conditions, some researchers concentrated on diffuse radiation modeling. Few examples as follows:
Jain (1990) derived several relations for estimating the global and diffuse radiations starting by expressing the intensities of direct and diffuse radiations as fractions of extraterrestrial radiation intensity. Two of the equations he derived were already known empirical equations including the Angstrom-Prescott equation. The other derived relations are:

\[
\frac{H_d}{H_0} = a_1 + b_1 \frac{H}{H_0} \quad (2.36)
\]

\[
\frac{H_d}{H} = a_2 + b_2 \frac{H_0}{H} \quad (2.37)
\]

\[
\frac{H}{H-H_d} = a_3 + b_3 \frac{S_0}{S} \quad (2.38)
\]

\[
\frac{H_d}{H-H_d} = a_4 + b_4 \frac{S_0}{S} \quad (2.39)
\]

\[
\frac{H}{H_d} = a_5 + b_5 \frac{S_0}{S} \quad (2.40)
\]

\[
\frac{H}{H_d} = A_1 + A_2 \frac{S_0}{S} + A_3 \left( \frac{S}{S_0} \right)^2 \quad (2.41)
\]

where \(H_d\) is the monthly average daily diffuse irradiation on a horizontal surface (J/m²).

The theoretical derivation showed that all the constants in the above equations are simple functions of three basic independent parameters. These parameters are the average transmission coefficient for diffuse radiation on horizontal surface for clear sky conditions, the average transmission coefficient for diffuse radiation on horizontal surface for cloudy sky conditions, and the average transmission coefficient for direct radiation for clear sky conditions.

Jain used the diffuse irradiation, the global irradiation and the bright sunshine duration data for Macerata (Italy), Salisbury and Bulawayo (Zimbabwe) to validate the derived models. The analysis showed good correlations for the linear equations.

Trabea (1999) investigated several empirical models in the literature that estimate the diffuse fraction radiation using sunshine duration or/and clearness index:
\[
\frac{H_d}{H} = a_1 + a_2 K \quad (2.42)
\]

\[
\frac{H_d}{H} = b_1 + b_2 K + b_3 K^2 \quad (2.43)
\]

\[
\frac{H_d}{H} = c_1 + c_2 \frac{S}{S_0} \quad (2.44)
\]

\[
\frac{H_d}{H} = d_1 + d_2 \frac{S}{S_0} + d_3 \left(\frac{S}{S_0}\right)^2 \quad (2.45)
\]

\[
\frac{H_d}{H_0} = e_1 + e_2 \frac{S}{S_0} + e_3 \left(\frac{S}{S_0}\right)^2 \quad (2.46)
\]

\[
\frac{H_d}{H} = f_1 + f_2 \frac{S}{S_0} + f_3 K \quad (2.47)
\]

To compare the performance of the above models, Trabea used global solar radiation, diffuse radiation and sunshine duration data, measured from 1982 to 1988 at different locations in Egypt to estimate the coefficients. These locations represent different weather conditions. He used the data of year 1992 to compare the performance of the models. Regression analysis showed that the best performance is for Equation (2.47) and the least is for Equation (2.46).

Ridley et al. (2010) developed a multi-variable logistic model for diffuse solar fraction (named BRL model). The BRL model uses hourly clearness index, daily clearness index, solar angle (solar elevation), apparent solar time (true solar time, which is based on the apparent motion of the actual Sun) and a measure of persistence of global radiation level as predictors:

\[
d = \frac{1}{1 + e^{-5.38 + 6.63k_t + 0.006AST - 0.007x + 1.75K_t + 1.31\Psi}} \quad (2.48)
\]

where \(d\) is hourly diffuse fraction, \(k_t\) is hourly clearness index, \(AST\) is apparent solar time, \(K_t\) is daily clearness index, and \(\Psi\) is a persistence factor of clearness index given by:

\[
\psi = \begin{cases} 
\frac{k_{t-1} + k_{t+1}}{2} & \text{sunrise} < t < \text{sunset} \\
k_{t+1} & t = \text{sunrise} \\
k_{t-1} & t = \text{sunset}
\end{cases} \quad (2.49)
\]
To build and validate the model they used data from Australian Bureau of Meteorology and different institutions in northern hemisphere. The analysis showed that the BRL model performs marginally better than currently used models for locations in the Northern Hemisphere and substantially better for Southern Hemisphere locations.

Furlan et al. (2012) developed a new regression model to estimate the hourly values of diffuse solar radiation at the surface. The model included the clearness index, the effects of cloud (cloudiness and cloud type), air temperature, relative humidity and atmospheric pressure at the surface and air pollution (concentration of particulate matter observed at the surface). They used the data of year 2002 for Pao Paulo, Brazil, since it contains complete records of the clouds. They applied a representative test to all meteorological variables and particulate matter concentrations. The test indicated that seasonal variations of these variables in 2002 were not statistically different from those based on 10 years of observation (1997–2006). To build the model, they used 75% of data and the rest to validate it. After building the model, they used variable ranking analysis to simplify the regression model. The analysis should that the meteorological variables and air pollution did not have any important effect, while the cloud information enhanced the model performance. Equation (2.50) shows the final simplified model.

\[ K_d = \beta_0 + \beta_1 (K - c)I_{K>0.228} + \beta_2 C + \beta_3 LowC + \beta_4 MiddleC + \beta_5 HighC \]  

where \( K_d \) is diffuse fraction \( \left( \frac{H_d}{H} \right) \), \( c \) is the break point \( (c = 0.228 \text{ for Pao Paulo, Brazil}) \) of the initial segmented model \( (K_d \sim K) \), \( I_{K>0.228} \) is an indicator function that assumed value 1 if \( (K > 0.228) \) and 0 otherwise, \( C \) is cloudiness \( (0 \text{ to } 10) \), \( LowC, MiddleC \) and \( HighC \) are factors assuming 1 and 0 to indicate, respectively, the presence and absence of low, middle and high clouds, and \( \beta' \text{s} \) are the parameters of the model.

Li et al. (2012) developed two models for general application in estimating the monthly average daily diffuse solar radiation in China. To build, validate and compare their models with four existing empirical models, they used data from 17 first-level meteorological stations across China, provided by the National Meteorological Information Centre. Their analysis showed that incorporating ambient temperature and relative humidity into empirical models could generally improve its estimates. The models are:
\[
\frac{H_d}{H} = 1.1937 - 0.6821K - 0.4658\frac{S}{S_0} - 0.0008T - 0.1987RH \tag{2.51}
\]

\[
\frac{H_d}{H_0} = 0.7537 - 0.5832\frac{S}{S_0} + 0.4954\log\frac{S}{S_0} - 0.0005T - 0.1123RH \tag{2.52}
\]

Magarreiro et al. (2014) reviewed solar radiation models that predict hourly diffuse fraction of global radiation. They divided the tested models into two categories. The first is diffuse-clearness index regression models, where these models are developed through piecewise fitting and divided into three intervals according to the hourly clearness index \((k_t)\) range. The first, second and third intervals represent data for overcast, partly cloudy and clear sky, respectively. The first and second \(k_t\) intervals correlations are usually linear and polynomial functions of \(k_t\), while the third interval has a constant value of diffusion fraction. A typical example is Miguel et al. (2001) model:

\[
d = 0.995 - 0.081k_t \quad \text{for} \quad k_t \leq 0.21 \tag{2.53-a}
\]

\[
d = 0.724 + 2.738k_t - 8.32k_t^2 + 4.967k_t^3 \quad \text{for} \quad 0.21 < k_t \leq 0.76 \tag{2.53-b}
\]

\[
d = 0.180 \quad \text{for} \quad k_t > 0.76 \tag{2.53-c}
\]

where \(d\) is the hourly diffuse fraction.

The second category includes diffuse fraction – clearness index and additional parameters of the regression models. The parameters used by different models include air mass, solar elevation, regional surface albedo, apparent solar time, a measure of persistence of global radiation and variability index (a diagnostic tool to detect statistically the presence of variable and inhomogeneous clouds).

The authors tested the applicability of all models to Azores Islands (mid latitude islands with typical Atlantic cloudy climate). They used Graciosa Island-Azores irradiance data, available in the Atmospheric Radiation Measurements (ARM) Climate Research Facility. In general, the models showed systematic underestimation of diffuse irradiance above 300 W/m\(^2\).

The following authors did similar type of work: Boland and Ridley (2008), Jacovides et al. (2010), Janjai et al. (2010), Karakoti et al. (2011), Li et al. (2011), Khalil and Shaffie (2013), Bortolini et al. (2013), Kuo et al. (2014).
CHAPTER 3: PROBLEM STATEMENT AND PROPOSED APPROACH

3.1 Approach

As mentioned in Chapter 1, global solar radiation data is very important for wide variety of applications and scientific studies. However, this data is not readily available because of the cost of measuring equipment and the tedious maintenance and calibration requirements. As a result, one major goal of the literature review presented in Chapter 2 is to find a model that can estimate and/or predict the global or diffuse radiation using other readily obtainable atmospheric parameters.

In this research, we have two major goals:

1- Interpretation: understand the coupling of global and diffuse solar radiation with climatic variability by revealing the relationships between these radiations and atmospheric parameters.
2- Prediction: build models that are able to capture, to a reasonable extent, the variability of global and diffuse radiation caused by their interactions with atmospheric parameters.

It is worthwhile mentioning that the interaction between solar radiation and the atmospheric parameters has physical basis (relationships). However, it is the complexity and dynamicity of the atmosphere that cause randomness in this process.

We notice that astronomical, geographical, and geometrical factors are not random; they can be determined accurately. In addition, their main effect is on extraterrestrial radiation amount. They are used to calculate extraterrestrial radiation at any time and location precisely (Duffie and Beckman, 2013). On the other hand, atmospheric parameters, which include the physical and meteorological factors, are random and they affect only the global radiation and its components. Accordingly, our approach is to:

1- Use visual multivariate analysis techniques to reveal, preliminarily, the relationship between solar radiations (global and diffuse) and the atmospheric parameters.
2- Use multiple linear regression (MLR) analysis and variable selection techniques to:
   a. Explore the relationship between global radiation and atmospheric parameters.
   b. Explore the relationship between diffuse radiation and atmospheric parameters.
3- Use different statistical learning techniques to build and validate models that are able to predict global or diffuse radiation from atmospheric parameters.

Statistical program R was used to perform all the analyses.

3.2 Data

This study is based on atmospheric parameters, global radiation and diffusion radiation data, in addition to the calculated extraterrestrial radiation. Fortunately, all these data are available in National Solar Radiation Database (NSRDB) website. The data of solar radiation provided by this website are a mix of measured and modeled data. However, only measured data were used in this research. This website has database for three periods:

1- NSRDB 1961-1990. This database includes 56 stations that have at least a portion of measured data of the 30-year record, and 183 stations with modeled data (Marion and Urban, 1995). The database of this period is not sufficient and suitable for this study and was not included in the analysis.

2- NSRDB 1991-2010 Update. This database holds solar and meteorological data for 1,454 locations in the United States and its territories. This database is the source of all the data used in the analysis of this research. The details are given in section 3.2.1.

3- NSRDB 2005-2012 Update. This database comprises 30-minute solar and meteorological data for approximately 1.4 million 0.038 ° latitude by 0.038 ° longitude surface pixels (nominally 4 km²), mostly modeled data.

3.2.1 Source of data

As mentioned before, the data source for this research is NSRDB 1991-2010 Update. Out of 1,455 locations, 38 stations have measured solar data, which were used in modeling the solar radiation for other sites. Figure 3.1 shows the sites of all locations with their classification, marking the stations with measured data. National Renewable Energy Laboratory (NREL) produced the figure in 2012. The stations’ classification is as follow (Wilcox, 2007):

Class I Stations have a complete period of record (all hours 1991–2010) for solar and key meteorological fields and have the highest-quality solar-modeled data (242 sites).
**Class II Stations** have a complete period of record but significant periods of interpolated, filled, or otherwise lower-quality input data for the solar models (618 sites).

**Class III Stations** have some gaps in the period of record but have at least 3 years of data that might be useful for some applications (594 sites).


Figure 3.1: NSRDB 1991-2010 stations with their classification, source NREL, April 2012

The data of twelve first class stations, which have measured data for both global and diffuse radiation, were chosen for this research; the details of these stations are in Table 3.1 and its keys. More details about the chosen stations and the quality of their data are in the Appendix (Wilcox, 2007).
Table 3.1: Class I stations that have both measured diffusion and global radiation.

<table>
<thead>
<tr>
<th>No.</th>
<th>ID</th>
<th>Station [Source]-State</th>
<th>No. of Years</th>
<th>Type of data</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>722140</td>
<td>Tallahassee Regional AP [ISIS] - FL</td>
<td>95-02 = 8</td>
<td>a</td>
</tr>
<tr>
<td>2</td>
<td>727640</td>
<td>Bismarck Municipal Arpt [ISIS] - ND</td>
<td>95-05 = 11</td>
<td>a</td>
</tr>
<tr>
<td>3</td>
<td>723650</td>
<td>Albuquerque Intl Arpt [ISIS] - NM</td>
<td>95-04 = 10</td>
<td>a</td>
</tr>
<tr>
<td>4</td>
<td>723870</td>
<td>Mercury Desert Rock AP [SURFRAD]-NV</td>
<td>98-05 = 8</td>
<td>a</td>
</tr>
<tr>
<td>5</td>
<td>726930</td>
<td>Eugene Mahlon Sweet Arpt [UO] - OR</td>
<td>96-05 = 10</td>
<td>a</td>
</tr>
<tr>
<td>6</td>
<td>722510</td>
<td>Corpus Christi Intl Arpt [UT] - TX</td>
<td>96-03 = 8</td>
<td>c</td>
</tr>
<tr>
<td>7</td>
<td>722540</td>
<td>Austin Mueller Municipal AP [UT]-TX</td>
<td>96-02 = 7</td>
<td>a</td>
</tr>
<tr>
<td>8</td>
<td>722660</td>
<td>Abilene Regional AP [UT] – TX</td>
<td>97-03 = 7</td>
<td>c</td>
</tr>
<tr>
<td>9</td>
<td>723630</td>
<td>Amarillo International AP [Canyon - UT] –TX</td>
<td>97-03 = 7</td>
<td>a</td>
</tr>
<tr>
<td>10</td>
<td>725720</td>
<td>Salt Lake City Int'l Arpt [ISIS] - UT</td>
<td>95-05 = 11</td>
<td>a</td>
</tr>
<tr>
<td>11</td>
<td>724030</td>
<td>Washington DC Dulles Int'l Ar [Sterling - ISIS] - VA</td>
<td>96-05 = 10</td>
<td>a</td>
</tr>
<tr>
<td>12</td>
<td>726410</td>
<td>Madison Dane Co Regional Arpt [ISIS] - WI</td>
<td>96-05 = 10</td>
<td>a</td>
</tr>
</tbody>
</table>

**Key1: data type**
- a – Measured global, direct, and diffuse with thermopile instruments
- b – Measured global and direct with thermopile instruments
- c – Spectrally corrected measured global and diffuse with rotating shadowband radiometer

**Key2: [Source of data]**
- UT: University of Texas Solar Energy Laboratory
- ISIS: Integrated Surface Irradiance Study network (National Oceanic and Atmospheric Administration (NOAA))
- SURFRAD: Surface Radiation Budget Measurement network (NOAA)
- UO: University of Oregon Solar Radiation Monitoring Laboratory Network

**Note:**
ID number is the six-digit United States Air Force (USAF) numbering system.

National Climatic Data Center (NCDC) provided all of the surface meteorological data for the entire period of record. While the measured solar radiation data were collected from several resources. For the stations included in the research, see Table 3.1 and its Key2. (Wilcox, 2007).
3.2.2 Data files

The data of research were extracted from Daily Statistical Files of NSRDB 1991-2010. Figure 3.2 is an example of these files for Albuquerque, New Mexico. The first group of data provides monthly and annual daily statistic for the period 1991-2010; following are the monthly daily statistics for each year. The details of Daily Statistical Files are in (Wilcox, 2007).

Figure 3.2: Daily Statistical File of NSRDB 1991-2010 for Albuquerque, New Mexico

The stations in Table 3.1 have a common period of data, from 1998 to 2002. These five years of data were used in the statistical analysis to explore the relationships between atmospheric parameters and solar radiations and as training data to build the models. In addition, the five underlined stations in Table 3.1 have a common period of data from 2003 to 2005. The data of this period were chosen as the test data used to evaluate the models performance.

Microsoft Excel was used to build appropriate training and test data files from the Daily Statistical Files of the stations in Table 3.1. The extracted data are the monthly average of daily values of the variables shown in Table 3.2. Each file of the training data contains the combined data of the twelve stations in Table 3.1. Each station has 60 (5 years × 12 months) rows of observations, where the row consists of the data of the 16 variables in Table 3.2. Accordingly, the
total number of rows in a data file is 720 (12 stations ×60 rows) and number of columns is 16 (one column for each variable). On the other hand, each file of the test data contains the combined data of the five underlined stations in Table 3.1. Each station has 36 (3 years×12 months) rows of observations for the 16 variables in Table 3.2. Accordingly, the test data file contains 180 (5 stations ×36 rows) rows and 16 columns (one column for each variable).

It is worthwhile mentioning that five of the stations included in the analysis are in relatively cold areas: one station each in North Dakota, Wisconsin, Utah, Oregon and Virginia. The other seven are in hot areas: one station each in Florida, New Mexico and Nevada, and four in Texas.

Table 3.2: Variables used in the analysis with their definitions and measure units

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$</td>
<td>Monthly average of total daily extraterrestrial solar radiation on a horizontal surface</td>
<td>Watthour/m$^2$</td>
</tr>
<tr>
<td>$H$</td>
<td>Monthly average of total daily global solar radiation on a horizontal surface</td>
<td>Watthour/m$^2$</td>
</tr>
<tr>
<td>$H_d$</td>
<td>Monthly average of total daily diffuse solar radiation on a horizontal surface</td>
<td>Watthour/m$^2$</td>
</tr>
<tr>
<td>$k$</td>
<td>Monthly average of clearness index ($H/H_0$)</td>
<td>unitless</td>
</tr>
<tr>
<td>$K_d$</td>
<td>Monthly average of diffuse fraction ($H_d/H$)</td>
<td>unitless</td>
</tr>
<tr>
<td>TotC</td>
<td>Monthly average of total sky cover</td>
<td>tenths</td>
</tr>
<tr>
<td>OpqC</td>
<td>Monthly average of Opaque sky cover</td>
<td>tenths</td>
</tr>
<tr>
<td>$H_2O$</td>
<td>Monthly average of precipitable water</td>
<td>cm</td>
</tr>
<tr>
<td>AOD</td>
<td>Monthly average of broadband aerosol optical depth</td>
<td>unitless</td>
</tr>
<tr>
<td>$T_{Max}$</td>
<td>Monthly average of maximum temperature</td>
<td>°C</td>
</tr>
<tr>
<td>$T_{Min}$</td>
<td>Monthly average of minimum temperature</td>
<td>°C</td>
</tr>
<tr>
<td>$T_{Range}$</td>
<td>Monthly average of temperature range</td>
<td>°C</td>
</tr>
<tr>
<td>$T$</td>
<td>Monthly average of temperature</td>
<td>°C</td>
</tr>
<tr>
<td>$T_{Daylight}$</td>
<td>Monthly average of daylight temperature</td>
<td>°C</td>
</tr>
<tr>
<td>RH</td>
<td>Monthly average of relative humidity</td>
<td>(%)</td>
</tr>
<tr>
<td>WS</td>
<td>Monthly average of wind speed</td>
<td>m/s</td>
</tr>
</tbody>
</table>
Table 3.2 shows the variables included in the analysis with their symbols, definitions and measurement units. For each month of a year, there are 16 variables. Only three of them are calculated, namely, clearness index $K$ ($K = \frac{H}{H_0}$), diffuse fraction $K_d$ ($K_d = \frac{H_d}{H}$) and temperature range $T_{Range}$ ($T_{Range} = T_{Max} - T_{Min}$).

For the total and opaque sky cover, NSRDB researchers have used the data of Automated Surface Observing System (ASOS) for clouds below 12,000 ft. and the ASOS Supplemental Cloud Product for clouds above 12,000 ft. For opaque cloud cover, researchers have used opacity factors of 1.00, 0.93, and 0.44 for low clouds, middle clouds and high clouds respectively, see Figure 3.3. The low cloud amounts and the middle and high cloud amounts are combined using an overlap equation to account for the low clouds overlapping the high and middle clouds:

$$sky\ cover\ (hundredths) = 100 - \frac{(100 - low\ amount) \times (100 - high\ and\ middle\ amount)}{100}$$

(3.1)

Figure 3.3: Pictorial description for cloud cover derivation (Wilcox, 2007)
3.2.3 Data testing

The goal of this testing is to examine the normality of the variables’ distributions. Figure 3.4 shows the normal QQ plot for the data of all variables in Table 3.2. The QQ plots in Figure 3.4 do not indicate profound deviations from normality.

Figure 3.4: Normal QQ plots for solar radiations and atmospheric parameters data
Figure 3.5 shows the probability distributions for the data of the variables in Table 3.2. These distributions are approximately normal. However, some of these distributions have two peaks. These two peaks are very clear in the distribution plots of global ($H$) and extraterrestrial ($H_0$) radiations. The existence of the two peaks is anticipated since the data were collected from cold regions and hot regions.
CHAPTER 4: CLOBAL RADIATION

This chapter studies the relationship between global radiation and atmospheric parameters.

4.1 Preliminary Analysis

Scatterplot matrix gives us preliminary assessment of causal relationships between plotted variables. Figure 4.1 shows the scatterplot matrix of global radiation and atmospheric parameters.

Figure 4.1: The scatterplot matrix of global radiation and atmospheric parameters
Figure 4.1 reveals different types of relationships between the variables. Plots, inside the green rectangle, show the relationship between K (fraction of solar radiation we receive at earth’s surface) and other variables, as well the relationship between H (global radiation) and other variables. No clear patterns exist between K or H and the variables H₂O (precipitation), AOD (aerosol optical depth), and WS (wind speed). In addition, Figure 4.1 shows a high collinearity between total cloud cover (TotC) and opaque cloud cover (OpqC). The high collinearity also exists among the variables of temperature, excluding TRange. Notice the plots inside the red rectangles.

The Correlation matrix of global radiation and atmospheric parameters, shown in Figure 4.2, confirms the above observations.

![Figure 4.2: The correlation matrix of global radiation and atmospheric parameters](image)

It is worthwhile mentioning that the correlation between any pair of variables, in the correlation matrix, does not extract the effect of other variables on that pair of variables. On the other hand, the coefficient of a predictor in a linear regression represents the marginal relationship between the response and that predictor, after removing the effect of other predictors (Faraway, 2002). Next, we used MLR to study the relationship between global radiation and atmospheric parameters.

### 4.2 Multiple Linear Regression Analysis

The maximum amount of solar radiation we can receive at earth’s surface is the extraterrestrial radiation (H₀), and that is in the absence of atmosphere. Since the extraterrestrial radiation is not constant (depends on time and location), the analysis included the clearness index $K$ as a response in some models. $K \left( K = \frac{H}{H_0} \right)$ represents the proportion of solar radiation
that we receive at earth’s surface. Three major MLR models with atmospheric parameters as predictors were analyzed to estimate the relationship between global radiation and atmospheric parameters.

4.2.1 Models with Clearness Index (K) response

The main goal of choosing K as a response is to study the effect of atmospheric parameters on the ratio of solar radiation that reaches the earth’s surface. All atmospheric parameters were included in the models as predictors. Although we expect to have a negative correlation between K and H₀, since \( K = \frac{H}{H_0} \), the correlation matrix in Figure 4.2 shows moderate positive correlation between them. Accordingly, H₀ was added to the predictors for inspection.

A- Variable selection

We used “best subset selection” technique, which gives us the best model based on the number of predictors. Then, we used “adjusting the training error for the model size” methods, such as adjusted R², Mallow’s Cₚ, and Bayesian information criterion (BIC), to determine the best model of all. Below is a brief explanation of these methods (James et al. 2014).

The best model is the one with the lowest test (data) mean squared error (MSE) given by Equation 4.1.

\[
MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \frac{RSS}{n} \quad (4.1)
\]

where \( n \) is number of observations, \( y_i \) is the \( i \)th response (observed value), \( \hat{y}_i \) is the \( i \)th fitted value and \( RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \) is residual sum of squares.

However, we first use training data to fit the models. Training MSE generally underestimates the test MSE. The reason is that we fit a model to the training data, using least squares, to estimate the regression coefficients such that the training RSS (but not the test RSS) is as small as possible. Therefore, training RSS cannot be used to select the best model among a set of models with different numbers of variables. To adjust for the training MSE any of the following approaches can be used:

\[
1- \quad C_p = MSE + \frac{2}{n} p \sigma^2 \quad (4.2)
\]
where \( p \) is the number of predictors, and \( \hat{\sigma}^2 \) is an estimate of the variance of the error \( \epsilon \) associated with each response measurement \( (\epsilon_i = y_i - \hat{y}_i) \). Note that \( \hat{\sigma}^2 = \frac{RSS}{n - p - 1} \).

Accordingly, the Cp statistic adds a penalty of \( \frac{2}{n} p \hat{\sigma}^2 \) to the training MSE in order to adjust for the fact that the training error tends to underestimate the test error. This penalty increases as the number of predictors in the model increases.

\[
2. \quad BIC = MSE + \frac{\log(n)}{n} p \hat{\sigma}^2 \quad (4.3)
\]

Since \( \log(n) > 2 \) for any \( n > 7 \), the BIC statistic generally places a heavier penalty on models with many variables. This results in the selection of smaller models than Cp does. Cp and BIC are indirect estimation of test MSE (James et al. 2014).

3. Adjusted \( R^2 \)

\( R^2 \) represents the proportion of variability in the responses \( Y (y_1 \ldots y_n) \) explained by the model:

\[
R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} \quad (4.4)
\]

where \( TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2 \) is the total sum of squares and it measures the total variability in the responses \( Y \), \( \bar{y} \) is the average of the responses \( Y \).

Since RSS always decreases as we add variables to the models, \( R^2 \) increases with this adding. Adjusted \( R^2 \) statistic, given by Equation 4.5, adds penalty on increasing the number of variables in a model.

\[
\text{Adjusted } R^2 = 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)} \quad (4.5)
\]

For Cp and BIC techniques, the best model is the model with the smallest value of Cp or BIC, while for adjusted \( R^2 \) the best model is the one with the largest adjusted \( R^2 \).

Applying “best subset selection” technique on the models with response \( K \), we obtained the results shown in Figure 4.3. We notice that for the one variable subset models, the best one is the opaque cloud cover (OpqC) model. This variable continues until the five variable subset model, after which, it is switched with total cloud cover (TotC). For the two variable subset,
relative humidity (RH) enters; it continues in the best subset models to the end. For the three variable subset, extraterrestrial radiation ($H_0$) enters and continues in the best subset models to the end.

![Table of selection algorithm]

To determine the best model of these best subset models, we applied “adjusting training error” criteria. Below is the adjusted $R^2$ values of the models shown in Figure 4.3, arranged in the same order. Figure 4.4 shows the selection results of adjusted $R^2$, Cp and BIC for the above models.

```
reg.summary$adj2
[1] 0.7178423 0.8342242 0.8672602 0.8727806 0.8798496 0.8871277 0.8904392
[8] 0.8935293 0.8948218 0.8954200 0.8956181
```

![Graphs of best models based on adjusted $R^2$, Cp and BIC]

Figure 4.3: Best subset selection results for the models with response K

Figure 4.4: Best model of K response subsets based on adjusted $R^2$, Cp and BIC selection
Figure 4.4 shows that the best model according to adjusted $R^2$ and Cp is the model with all variables, while BIC selected the model with nine variables. However, after adding the third variable, the improvement in test MSE starts to flatten. For convenience and simplicity, we analyzed the four variable model.

**B- Four variable model**

The best model of the four variable subset is the model with the predictors $OpqC$, RH, $H_0$, and $T_{Range}$.

**Fitting the model**

Figure 4.5 shows the results of fitting the four variable model. The coefficients of the model are very significant for all variables. $K$ has negative correlation with $OpqC$ and RH, and positive correlation with $H_0$ and $T_{Range}$. The coefficient of $H_0$ is very small compared to other coefficients. This model explained only 87.28% of the variability in $K$ responses. It is worth mentioning that $0.2 < K < 0.8$. The typical value of $K$ for clear sky day is between 0.65 and 0.75 (Suehrcke, 2000).

![Call output](image)

Figure 4.5: The results of fitting $K$ against $OpqC$, RH, $H_0$, and $T_{Range}$ variables

Equation 4.6 represents the fitted model:

$$K = 0.701 + 6.379 \times 10^{-6}H_0 - 0.023OpqC - 0.002RH + 0.004T_{Range} + \epsilon$$  \hspace{1cm} (4.6)
Diagnostic analysis

Figure 4.6 shows four major diagnostic plots for model (4.6). The residuals plot, A, shows no profound pattern of nonlinearity of data or non-constant variance of residuals, although there is a small accumulation of points below the horizontal line $\epsilon = 0$, at the right corner. These points belong to NV station for June, July and August data.

![Residuals vs Fitted](image1)

![Normal Q-Q Plot](image2)

![Jackknife Residuals vs Fitted](image3)

![Jackknife Residuals vs Leverage](image4)

Figure 4.6: Diagnostic plots for model (4.6). A- Linearity and constant variance of $\epsilon$ test. B- Normality test ($\epsilon \sim N(0, \sigma^2)$). C- Outliers test. D- High leverage points test.

Normal Q-Q plot, B, tests the assumption $\epsilon \sim N(0, \sigma^2)$. The plot shows no significant deviation from normality, where the majority of the points lie on the Q-Q line. The point above
the Q-Q line represents UT Jan-2001. While the points below the Q-Q line represent (starting from the furthest point) ND Jan-1999, WI Jan-1999, WI Jan-2000, and WI-Jan-2001. All these points are potential outliers.

Studentized residuals were used in Q-Q plot, as well in plots C and D. Studentized residuals, also called Jacknife residuals, are calculated using Equation 4.7:

\[ \text{Studentized } \epsilon_i = \hat{y}(i) - y_i \]  

\[ (4.7) \]

where \( \hat{y}(i) \) is the predicted value of the response \( i \), calculated form a model fitted by excluding the point \( i \); \( y_i \) is the \( i \)th observed response.

Accordingly, Studentized residuals reveal the possible outliers, which pull the regression line so close to them that they conceal their true status. If the Studentized residual of a point is large, then this point is an outlier. The red lines in Figure 4.6-C represent the Bonferroni critical value of Studentized residuals, beyond which the points are outliers (Faraway, 2002). The critical value calculated using Equation 4.8 with \( \alpha = 0.05 \).

Bonferroni critical value = quantile of \( t \left( \frac{\alpha}{2n}, n - p - 1 \right) \)  

\[ (4.8) \]

We notice that UT Jan-2001 is an outlier; ND Jan-1999 and WI Jan-1999 are almost outliers. These points are mild outliers; they deviate slightly from the Q-Q line in plot B. Given that we have 720 points, these outliers are of no concern, especially, they do not have high-leverage. Figure 4.6-D reveals the high-leverage points, the red line (Leverage = 2p/n) in plot D is just a “rule of thumb” critical value. Two points have serious high-leverage, Corpus Christi-TX Jun-1999 and Jul-1999. Fortunately, they are not outliers and consequently not influential points.

An important assumption of the linear regression model is that the error terms, \( \epsilon_1, ..., \epsilon_n \) are uncorrelated. If there is correlation among error terms, the true standard errors will be underestimated. In this case, confidence and prediction intervals will be narrower than they should be, and p-values associated with model will be lower than they should be (James et al. 2014). Figure 4.7 shows residuals verses time for two stations, Tallahassee Regional AP –FL on the left and Corpus Christi Intl Arpt –TX on the right. No profound correlation exists among the residuals.
Finally, we computed the variance inflation factor (VIF) of the model variables to check for multicollinearity problem. Denoting the estimated value of the model coefficient for variable $j$ by $\hat{\beta}_j$, the VIF is the ratio of the variance of $\hat{\beta}_j$ when fitting the full model divided by the variance of $\hat{\beta}_j$ if fit on its own. The smallest value for VIF is one, which indicates a complete absence of collinearity. As a rule of thumb, a VIF value that exceeds 5 or 10 indicates a problematic amount of collinearity (James et al, 2014). VIF values for the model (4.6) predictors are VIF ($H_0$) = 1.15, VIF (OpqC) = 2.49, VIF (RH) = 1.89, and VIF (T_{\text{Range}}) = 2.27. These values indicate that there is no collinearity problem.

**Comparison with other models**

*Five variable model:* The best model includes $H_0$, OpqC, RH, T, and T_{Daylight}. We calculated the VIF values of this model to test for the collinearity. VIF values are VIF ($H_0$) = 3.49, VIF
(OpqC) = 3.03, VIF (RH) = 1.83, VIF (T_{Daylight}) = 434.99, and VIF (T) = 441.18. The last two values indicate a big collinearity problem. Accordingly, the four variables model is preferred.

*Model without predictor H_0:* Since the model coefficient of variable H_0 is very small, we repeated the analysis excluding H_0. The best “one variable model” is again OpqC model. For The two variable model, RH enters. Both variables continue to the end of best subsets. For the three variable model, T_{Daylight} enters. In four variable model, T_{Daylight} leaves and T_{Min} and T_{Range} enter, where both variables continue to the end of best subsets. However, the adjusted R^2 values are slightly better for the model including H_0, starting from the three variable model, when H_0 enters the model. The comparison is below:

Adjusted R^2 for the model with H_0 predictor:

```
> reg.summary$adjr2
[1] 0.7178423 0.8342242 0.8672602 0.8727806 0.8798496 0.8871277 0.8904392
[8] 0.8935293 0.8948218 0.8954200 0.8956181
```

Adjusted R^2 for the model without H_0 predictor:

```
> reg.summary$adjr2
[1] 0.7178423 0.8342242 0.8620772 0.8708668 0.8766575 0.8800515 0.8809933
[8] 0.8818989 0.8823969 0.8825050
```

Equation 4.9 represents the fitted four variables model without H_0:

\[
K = 0.7094 + 0.0018T_{Min} - 0.0204OpqC - 0.0025RH + .0062T_{Range} + \epsilon \\
(4.9)
\]

The diagnostic analysis gave similar results.

4.2.2 Models with log (H) response

We noticed that model (4.6) explained only 87.3% of the variability of clearness index (K). To enhance the model performance, we used log (K). However, we chose log (H) to be the response based on the flowing equation.

\[
log(K) = log(H/H_0) = log(H) - log(H_0) \\
(4.10)
\]

We transferred log (H_0) to the predictors’ side and set it as an offset predictor. In this case, the fitted model represents the effect of atmospheric parameters on log (K) and enable us to predict H easily.
A- Variable selection

Applying best subset selection technique excluding log (H₀) made the temperature variables, which are highly correlated, the most three important variables and lowered the adjusted R² values. Accordingly, log (H₀) treated as a predictor. Figure 4.8 shows the results of best subset selection technique, for the models with log (H) response.

![Table of best subset selection results](image)

Figure 4.8: Best subset selection results for the models with log (H) response

As expected, log (H₀) is the most important variable and it continues in the best subset models to the end. For the two variable model, OpqC enters and continues to the end. For the three variable model, RH enters and continues to the end. Below is the adjusted R² values for the models in Figure 4.8. Figure 4.9 shows the best model according to adjusted R², Cp, and BIC.

```
reg.summary$adjr2
[1] 0.9997279 0.9674916 0.9787768 0.9796392 0.9807179 0.9819036 0.9822842 0.9826987
[9] 0.9828883 0.9829085 0.9828847
```

![Graphs of adjusted R², Cp, and BIC](image)

Figure 4.9: Best log (H) response model of best subsets based on adjusted R², Cp and BIC selection
We notice that adjusted $R^2$ improved when we used log (H) as a response. For four variable model, adjusted $R^2$ is 0.98, while for K response model this value is 0.87. Figure 4.9 shows that the improvement in the log (H) response models flatten after the third variable. For convenience, four variable model was analyzed.

B- Four variable model

The best model of the four variable subset is the model with the predictors OpqC, log ($H_0$), RH, and $T_{Range}$.

Fitting the model

Figure 4.10 shows the results of fitting the log (H) response model with four variables. The coefficients of all variables are very significant. Log ($H_0$) has the largest coefficient. Log (H) has positive correlation with log ($H_0$) and $T_{Range}$, while it has negative correlation with OpqC and RH.

![Figure 4.10: The results of fitting log (H) against log ($H_0$), OpqC, RH, and $T_{Range}$](image)

Equation 4.11 represents the fitted model

$$\log(H) = -1.212 + 1.104 \log(H_0) - 0.045OpqC - 0.004RH + .009T_{Range} + \epsilon$$ (4.11)
Diagnostic analysis.

Figure 4.11 shows the diagnostic plots of model (4.11). The features are similar to those of model (4.6). Plot (A) shows no profound pattern of nonlinearity of data or non-constant variance of residuals. Plot (B) shows no significant deviation from normality, where the majority of the points lie on the Q-Q line. Plot (C) shows the same outliers as in Figure 4.6-C with extra outlier points. The outlier points belong to UT Jan-01, WI Jan-01, OR Jan-98, and ND Jan-99. Plot (D) shows the same pattern of high-leverage points as in model (4.6).

```
\begin{figure}[h]
\centering
\includegraphics[width=\linewidth]{fig4_11.png}
\caption{Diagnostic plots for model (4.11). A- Linearity and constant variance of \( \epsilon \) test. B- Normality test \( \epsilon \sim N(0, \sigma^2) \). C- Outliers test. D- High leverage points test.}
\end{figure}
```

“The correlation of residuals” plot for model (4.11) is the same as in Figure 4.7. Accordingly, it is not displayed. There is no profound correlation among the residuals. Finally, the VIF values of model (4.11) are VIF (log (H0)) = 1.17, VIF (OpqC) = 2.48, VIF (RH) = 1.89, and VIF (TRange) = 2.30. These values indicate that there is no significant collinearity problem among variables.
Comparison with other models

Five variable model: The best model includes log \((H_0)\), OpqC, RH, T, and \(T_{\text{Daylight}}\). The calculated VIF values of this model are VIF \((\log(H_0)) = 3.29\), VIF \((\text{OpqC}) = 3.01\), VIF \((\text{RH}) = 1.83\), VIF \((\text{T}) = 403.19\), and VIF \((\text{T}_{\text{Daylight}}) = 404.84\). The last two values indicate a big collinearity problem. Accordingly, the four variable model is preferred.

Model with \(\log(H_0)\) as an offset predictor: Based on Equation (4.10), to study the effect of atmospheric parameters on the response \(\log(K)\), we need to offset the predictor \(\log(H_0)\). We refitted model (4.11), making \(\log(H_0)\) an offset predictor. Figure 4.12 shows the results of the new fitted model.

Equation 4.12 represents the fitted model.

\[
\log(H) = -0.320 + \log(H_0) - 0.049OpqC - 0.004RH + .012T_{\text{Range}} + \epsilon \quad (4.12)
\]

We notice that the adjusted \(R^2\) became 0.9712 compared to 0.9796 for model (4.11). The decrease in adjusted \(R^2\) is small and the two models are close.

Model with zero intercept and offset \(\log(H_0)\): Fitting model (4.12) with zero intercept caused small degradation in the model, where residual standard error increased from 0.07647 to 0.08171, that is 7% increase. Models with smaller residual standard error have better performance.
4.2.3 Models with H response

In the third group of models, we studied the effect of atmospheric parameters on the amount of global radiation that reaches the earth’s surface, where H was used directly as a response.

A- Variable selection

Applying best subset selection technique for H response models, we obtained the results shown in Figure 4.13. The first variable to enter is $H_0$ and it continues to the end. For the two variable model, RH enters and stays until the end. The third variable to enter is the total cloud cover (TotC) and it continues to the end. Notice that for the models related to the clearness index response (K), opaque cloud cover entered not TotC. The fourth variable to enter is the average temperature (T) and it continues to the end.

![Figure 4.13: Best subset selection results for the models with H response](image)

Below is the adjusted $R^2$ values for the models in Figure 4.13, arranged in the same order. We notice that the adjusted $R^2$ values are large. Figure 4.14 shows the best model of the subsets according to adjusted $R^2$, Cp, and BIC.

```
reg.summary$adjr2
[1] 0.8764539 0.9626362 0.9756199 0.9760605 0.9773200 0.9773858 0.9781812
[8] 0.9782192 0.9782588 0.9782630 0.9782425
```
Figure 4.14 shows that the improvement in H response models flatten after the third variable, in the same manner as the previous models did. Four variable model was analyzed.

**B- Four variable model**

The best model of the four variable subset is the model with the predictors $H_0$, RH, TotC, and T.

**Fitting the model**

Figure 4.15 shows the results of fitting the four variable model of H response. All coefficients are highly significant. The correlation pattern is the same as the previous models.

```
Call:
lm(formula = H ~ H0 + TotC + RH + T, data = MData44Var)

Residuals:
    Min     1Q  Median     3Q    Max
-1306.91 -166.50   -3.41  193.53  817.83

Coefficients:               Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.005e+03   6.290e+01   31.917  < 2e-16 ***
H0           5.831e-01    7.308e-03    79.786  < 2e-16 ***
TotC         -1.522e+02    8.837e+00   -17.222  < 2e-16 ***
RH           -2.409e+01    8.463e-01   -28.462  < 2e-16 ***
T            8.035e+00    2.134e+00     3.765  0.00018 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1

Residual standard error: 289.3 on 715 degrees of freedom
Multiple R-squared:  0.9762, Adjusted R-squared:  0.9761
F-statistic: 7330 on 4 and 715 DF,  p-value: < 2.2e-16
```

Figure 4.15: The results of fitting H against $H_0$, TotC, RH, and T
Equation 4.13 represents the fitted model

$$H = 2.008 \times 10^3 + 0.583H_0 - 1.522 \times 10^2TotC - 24.09RH + 8.035T + \epsilon \quad (4.13)$$

Diagnostic analysis

Figure 4.16 shows the diagnostic plots of model (4.13). Plot (A) shows no profound pattern of nonlinearity of data or non-constant variance of residuals, although there is a small upward curvature. Plot (B) shows no significant deviation from normality. Plot (C) shows one outlier that belongs to OR May-98. Plot (D) shows no problems caused by high-leverage points.

Figure 4.16: Diagnostic plots for model (4.13). A- Linearity and constant variance of $\epsilon$ test. B- Normality test ($\epsilon \sim N(0, \sigma^2$). C- Outliers test. D- High leverage points test.

Figure 4.17 shows the correlation of residuals plot for model (4.13). There is no profound correlation among the residuals.
Finally, the VIF values of model (4.13) variables are VIF ($H_0$) = 3.13, VIF (TotC) = 2.00, VIF (RH) = 1.73, and VIF (T) = 3.44. These values indicate that there is no significant collinearity problem among the variables.

**Comparison with other models**

*Five variable model:* The best model includes $H_0$, TotC, RH, T, and $H_2O$ parameters. We noticed that the five variable model, in the K response and log (H) response, contains the highly correlated variables T and $T_{Daylight}$. In H response model, the fifth variable is precipitation ($H_2O$). The calculated VIF values of this model are VIF ($H_0$) = 3.13, VIF (TotC) = 2.07, VIF (RH) = 3.73, VIF (T) = 8.11, and VIF ($H_2O$) = 6.33. The last two values indicate a small to moderate collinearity problem. Accordingly, the four variable model is preferred. However, the five variable model was included in prediction analysis. Equation 4.14 represents this model.

\[
H = 1780.76 + 0.58H_0 - 162.42TotC - 18.44RH + 23.46T - 160.58H_2O + \epsilon \quad (4.14)
\]

All the model coefficients are very significant.

Figure 4.17: Residuals vs Time for H response model. Left side is for FL station data, Right side is for Corpus-TX station.
4.3 Prediction Analysis

In the previous section, we fitted four models with response H or log (H). These models are:

\[
\log(H) = -1.212 + 1.104 \log(H_0) - 0.045OpqC - 0.004RH + .009T_{Range} + \epsilon \quad (4.11)
\]

\[
\log(H) = -0.320 + \log(H_0) - 0.049OpqC - 0.004RH + .012T_{Range} + \epsilon \quad (4.12)
\]

\[
H = 2.008 \times 10^3 + 0.583H_0 - 1.522 \times 10^2TotC - 24.09RH + 8.035T + \epsilon \quad (4.13)
\]

\[
H = 1780.76 + 0.58H_0 - 162.42TotC - 18.44RH + 23.46T - 160.58H_2O + \epsilon \quad (4.14)
\]

In this section, we tested the ability of these models to predict the global radiation, given the atmospheric parameters shown in Equations 4.11- 4.14. It is worth confirming that, although H response models and log (H) response models were used to predict global radiation (H), they have different interpretations. The predictors in log (H) response models represent the factors that most affect the ratio of solar radiation that we receive at earth’s surface (see Equation 4.10). On the other hand, the predictors in H response models represent the factors that most affect the amount of global radiation received at earth’s surface.

The accuracy and performance of the models were assessed using the following statistical indicators:

1- The adjusted coefficient of determination (adjusted R^2), defined in Equation 4.5
2- The mean percentage error (MPE):

\[
MPE = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{\hat{y}_i - y_i}{y_i} \right) \times 100 \quad (4.15)
\]

3- The mean absolute percentage error (MAPE):

\[
MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{\hat{y}_i - y_i}{y_i} \right| \times 100 \quad (4.16)
\]

4- The root mean squared error (RMSE):

\[
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \quad (4.17)
\]
5- The mean bias error (MBE):
\[
MBE = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)
\]  \hspace{1cm} (4.18)

As mentioned before, the adjusted R\(^2\) value represents the percentage or fraction of variation in the response explained by the model. The largest value of adjusted R\(^2\) is one. MPE can be defined as the percentage deviation between predicted and measured monthly average daily global radiation data. MAPE is the average absolute value of percentage deviation between predicted and measured global radiations. MBE reveals whether a given model has a tendency to underestimate or overestimate the global radiation in the long term. RMSE indicates the level of scatter that a model produces by providing a term-by-term comparison of the actual deviation between the predicted and measured global radiation values. Low values of MPE, MAPE, RMSE, and MBE indicate a good performance (Ertekin and Yaldiz, 2000; Ertekin and Evrendilek, 2007; Yao et al., 2014).

The data used to test the models performance belong to the stations shown in Table 4.1.

Table 4.1: The test data used to evaluate models performance.

<table>
<thead>
<tr>
<th>No.</th>
<th>ID</th>
<th>Station [Source]-State</th>
<th>Years</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>723650</td>
<td>Albuquerque Intl Arpt [ISIS] - NM</td>
<td>2003-2005</td>
</tr>
<tr>
<td>2</td>
<td>723870</td>
<td>Mercury Desert Rock AP [SURFRAD]-NV</td>
<td>2003-2005</td>
</tr>
<tr>
<td>4</td>
<td>725720</td>
<td>Salt Lake City Int'l Arpt [ISIS] - UT</td>
<td>2003-2005</td>
</tr>
<tr>
<td>5</td>
<td>727640</td>
<td>Bismarck Municipal Arpt [ISIS] - ND</td>
<td>2003-2005</td>
</tr>
</tbody>
</table>

The data of these stations, from 1998 to 2002, was part of the data used in fitting the global radiation models. However, to test the models performance, we need different data. Accordingly, we used the data of years 2003-2005. The test data file contains 180 (5 stations × 3 years × 12 months) rows of observations with 16 columns (one for each variable). Two of the chosen stations represent hot climate (NM and NV) and two represent cold climate (UT and ND). While Virginia represents moderate climate. Since the test data file for years 2003-2005 of the five stations contains 180 vectors of observations (y, x\(_1\), x\(_2\)…x\(_p\)), the quantity of test data is 25% of the training data quantity.
Table 4.2 shows the validation results of H and log (H) response models using the test data.

Table 4.2: The accuracy of global models based on the statistical error tests.

<table>
<thead>
<tr>
<th>Model</th>
<th>MPE</th>
<th>MAPE</th>
<th>RMSE</th>
<th>MBE</th>
<th>Adjusted R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>(4.11) log (H)</td>
<td>-1.005%</td>
<td>5.925%</td>
<td>327.194</td>
<td>-26.020</td>
<td>0.973</td>
</tr>
<tr>
<td>(4.12) log (H) - offset (log(H₀))</td>
<td>-0.915%</td>
<td>5.977%</td>
<td>330.091</td>
<td>-82.934</td>
<td>0.9725</td>
</tr>
<tr>
<td>(4.13) H – 4 predictors</td>
<td>-0.155%</td>
<td>5.917%</td>
<td>308.529</td>
<td>22.469</td>
<td>0.976</td>
</tr>
<tr>
<td>(4.14) H – 5 predictors</td>
<td>-1.119%</td>
<td>6.378%</td>
<td>303.934</td>
<td>8.113</td>
<td>0.977</td>
</tr>
</tbody>
</table>

We mentioned in section 4.2 that the best model is the one with the lowest test (data) mean square error (MSE) given by Equation 4.1. Consequently, in Table 4.2, the major statistical indicator is RMSE. From Table 4.2, we notice that the RMSE values of H response models (models (4.13) and (4.14)) are smaller than the RMSE values of log (H) response models (models (4.11) and (4.12)). This indicates that the prediction ability of H response models is better than the prediction ability of log (H) response models. The lowest value of RMSE is for model (4.14) (H response model with five predictors), accordingly this model has the best performance. However, there is no significant difference between the performances of H response models. The same applies for log (H) response models, where their performances are very close. Finally, the values of RMSE obtained for the models in Table 4.2 are similar and close to those obtained in the literature. For example, converting the units of RMSE to MJ/m² for model (4.13), we obtain 1.1MJ/m². This value is considered good and acceptable.

The indicator adjusted R² supports the above results. The values of adjusted R² indicate that H response models perform better than log (H) response models, and the best performance is for model (4.14). In addition, the adjusted R² values for log (H) response models are almost the same, and the adjusted R² values for H response models are very close. We notice that the adjusted R² values for all the models are above 0.97. This means that those models are able to explain at least 97% of the variability of global radiation (H). However, it is worth mentioning that, the training data gave us adjusted R² values for log (H) response models higher than the adjusted R² values of H response model. Accordingly, we expected log (H) response models to perform better than H response models, but the opposite happened. This could happen since the difference
between adjusted $R^2$ values for H and log (H) response models are very small (e.g. 0.003 for three variable models).

Log (H) response models have negative values of the mean bias error (MBE). This indicates that log (H) response models tend to underestimate the global radiation (predicted values are smaller than measured values). On the other hand, H response models have positive values of MBE. This indicates that H response models tend to overestimate the global radiation (predicted values are larger than measured values).

The values of the mean percentage error (MPE) of all the models are negative and very small. These MPE values just indicate that the percentage error of underestimated values slightly prevails the percentage error of overestimated values. However, a drawback of MPE, as well of MBE, that an overestimation of a response can be hidden by an underestimation of another response. Consequently, the mean absolute percentage error (MAPE) is a more realistic estimation of the average percentage deviation between predicted and measured values of global radiation. Table 4.2 shows that the MAPE values of all the models are around 6%. This value is reasonable and accepted.

It is worth mentioning that the comparison among the models performance was based on the predicted and measured values of H. This includes the log (H) response models, where the predicted values of log (H) were converted to H. The reason is that we want to test the ability of the models to predict global radiation H not log (H).

**4.4 Discussion and Conclusions**

Using MLR analysis to understand the interaction between global radiation and atmospheric parameters, revealed the following:

1- Relative humidity (RH) and cloud cover are the main atmospheric parameters that affect the proportion of solar radiation (K) and amount of global radiation (H) we receive at earth’s surface. K and H have negative correlation with cloud cover. This is expected, because of the cloud cover role in absorbing, reflecting back and scattering the solar radiation, which causes a reduction in the global radiation we receive at earth’s surface. K and H have also negative correlation with RH. This could be attributed to the effect of
water vapor molecules and small water drops in the atmosphere, which absorb, reflect and scatter the photons.

2- Although \( K = \frac{H}{H_0} \), it has positive correlation with extraterrestrial radiation \((H_0)\). This means that as extraterrestrial radiation increases, the proportion of solar radiation we receive at earth’s surface increases. This is reasonable for the United States climates, where in summer times (large amounts of \(H_0\)) the frequency of cloud cover occurrence becomes less and consequently we have more direct and global radiation. However, the regression coefficient of \(H_0\) is very small for the K response models. \(H\) also has positive correlation with \(H_0\). This is natural since \(H_0\) represents the original amount of solar radiation, before the atmosphere dissipates part of it.

3- The importance of atmospheric parameter effect depends on the response variable of the model. Table 4.3 shows the effective parameters, in descending importance, for the major global radiation models discussed in section 4.2

<table>
<thead>
<tr>
<th>Order</th>
<th>Model response</th>
</tr>
</thead>
<tbody>
<tr>
<td>K</td>
<td>Log (H)</td>
</tr>
<tr>
<td>First</td>
<td>OpqC (opaque cloud cover)</td>
</tr>
<tr>
<td>Second</td>
<td>RH</td>
</tr>
<tr>
<td>Third</td>
<td>(H_0)</td>
</tr>
<tr>
<td>Fourth</td>
<td>(T_{Range})</td>
</tr>
</tbody>
</table>

From Table 4.3, we notice that for the models related to the proportion response, such as \(K\) and log \((H)\), opaque cloud cover and \(T_{Range}\) are effective variables; while for the amount response \((H)\), total cloud cover and average temperature \((T)\) are the effective. Since the correlation or collinearity between opaque cloud cover and total cloud cover is very high, replacing one variable with the other will cause negligible effect on the model performance. For example, replacing \(TotC\) by \(OpqC\) in the \(H\) response model (4.13) reduced adjusted \(R^2\) by 0.008 only. For log \((H)\) response model (4.11), replacing \(OpqC\) by \(TotC\) reduced adjusted \(R^2\) by 0.005 only. Both variables are very significant when exist in the model separately. In contrast, the correlation between \(T_{Range}\) and average temperature
(T) is small (0.306). Though, replacing one variable with the other has negligible effect on the model performance. For example, replacing T by TRange in the H response model (4.13) reduced adjusted R² by 0.003 only. For log (H) response model (4.11), replacing TRange by T reduced adjusted R² by 0.008 only. The reason is that both variables enter in the four variable model, where the improvement in models performance, as we saw in variable selection analysis, start to flatten considerably. However, T is insignificant predictor for log (H) model (4.11), where it has very small t- value (-0.414) with probability 0.679. In addition, TRange significance is less than T significance in H response model (4.13). It is worth mentioning that TRange can be considered as a surrogate or indicator to the station site. When TRange values for a station are large, this indicates that the station is located in a dry or clear sky region. On the other hand, small values of TRange indicates that the station is located in a wet or cloudy region. For the models with H or log (H) response, the most effective parameter is \( H_0 \) or log \( (H_0) \). This is natural since \( H_0 \) represents the maximum amount of solar radiation we can receive theoretically (in the absence of atmosphere). For the K response model, OpqC is the most effective variable, where the cloud cover can reduce the proportion of solar radiation received on earth’s surface considerably by reflection, absorption and scattering. Temperature variable occupies the forth place in all models, whether it is TRange or average temperature (T). Unexpectedly, RH occupies the second place and TotC occupies the third place in H response models.

4- Adding wind speed (WS), aerosol optical depth (AOD), and water precipitation (H₂O) variables to the model predictors has no significant improvement on the model performance. This indicates that their effect on global radiation is negligible.

5- After the forth predictor, we start to have collinearity problem by adding more variables. The reason is the high collinearity among the temperature variables and between total and opaque cloud covers. Due to the collinearity problem, adding more variables to the model did not cause significant improvement in the model performance. We also noticed moderate collinearity between average temperature (T) and water precipitation (H₂O) in the five variable model of H response. As a result, adding H₂O predictor to the model did not improve the model performance significantly.

6- The three parameters H₀, RH and cloud cover were able to explain around 87 % of the variability in K response, and around 98 % of the variability in H and log (H) responses.
Accordingly, having data for \( H_0 \), RH, and TotC, which can be easily obtained, will enable us to predict monthly average of total daily global radiation efficiently, using the following simple model.

\[
\hat{H} = 1.957 \times 10^3 + 0.605H_0 - 1.637 \times 10^2TotC - 2.347 \times 10RH \quad (5.19)
\]

where \( \hat{H} \) is the predicted value of monthly average of daily global radiation.

7- The validation test for \( H \) and log (\( H \)) response models showed that the models were able to predict the global radiation effectively, with a superiority for \( H \) response models. However, \( H \) response models tend to overestimate global radiation and log (\( H \)) response models tend to underestimate global radiation. Although, the best performance is for model (4.14) (the \( H \) response model with five predictors), model (4.13) (\( H \) response model with four predictors) is preferred since it is more simple and its performance is very close to model (4.14) performance.

Our work is similar to Ertekin and Yaldiz (1999) work. The authors used multiple linear regression models to estimate the monthly average daily global radiation for Antalya, Turkey. They used nine variables: \( H_0 \), RH, T, \( H_2O \), TotC, solar declination, sunshine duration, soil temperature, and evaporation. They concluded that the best model is the nine variable model based on correlation coefficient value of 0.9986. In our analysis we concentrated on identifying the atmospheric parameters that physically and substantially affect the amount of global radiation, not a result or an indication of it such as sunshine duration or soil temperature. The objective was to fit simple models that able to capture the effect of climate variability on global radiation and predict it effectively. The performance of our simple models, which have only \( H_0 \), RH and TotC predictors (adjusted \( R^2 \) values of 0.98) is comparable to the performance of Ertekin and Yaldiz models.
CHAPTER 5: DIFFUSE RADIATION

The component of global radiation that most depends on the atmospheric conditions is the diffuse radiation. Consequently, the focus in this chapter was on diffuse radiation. The diffuse radiation analysis examined two responses, the monthly average of daily diffuse radiation on a horizontal surface \( H_d \) and monthly average of daily diffuse fraction \( K_d = \frac{H_d}{H} \). We studied the dependence of these two responses on atmospheric parameters.

5.1 Preliminary Analysis

Figure 5.1 shows the scatterplots and the correlations of diffuse fraction \( K_d \) and diffuse radiation \( H_d \) with atmospheric parameters, clearness index, and global radiation. We notice that the correlation between clearness index \( K \) and diffuse fraction \( K_d \) is very high (-0.905). On the other hand, although we expected a high correlation between \( K_d \) and \( H_d \), the correlation between them is low (0.153). In addition, they correlate differently with atmospheric parameters. However, as mentioned before, the scatterplot matrix and correlation matrix do not exclude the effect of other parameters on the pairs of variables. The correlation between a pair of variables, after extracting the effect of other parameters, is represented by the coefficients of regression model, where one of the variables is the response.

![Figure 5.1: The scatterplots and correlations of average daily diffuse radiation and daily diffuse fraction with atmospheric parameters](image)

5.2 Multiple Linear Regression Analysis

We studied the effect of atmospheric parameters on the amount of diffuse radiation \( H_d \) and on diffuse fraction \( K_d \) using MLR.
5.2.1 Models with Diffuse Fraction ($K_d$) Response

We noticed the high correlation between $K \left( K = \frac{H}{H_0} \right)$ and $K_d \left( K_d = \frac{H_d}{H} \right)$. As a result, many articles in the literature investigated this relationship to predict the $K_d$ from $K$. In this study, two parallel analyses were conducted. The first included $K$ as a predictor, and the second excluded it to concentrate on the effect of atmospheric parameters on $K_d$.

Variables selection

1- Models with clearness index ($K$) predictor.

Figure 5.2 shows the results of “best subset selection” technique for the $K_d$ response models. As expected, $K$ is the first variable to enter and it continues to the end. For the two variable model, aerosol optical depth (AOD) enters and it continues to the end. For the three variable model, RH enters and continues to the end. In the four variable model, OpqC enters and it leaves in the five variable model, where $H_2O$ and $H_0$ enter and continue to the end.

To determine the best model of these best subset models, we applied “adjusting training error” techniques. Below is the adjusted $R^2$ values of the models shown in Figure 5.2, arranged in the same order. We notice that $K$ explains about 82% of the variability in $K_d$ response. The improvement in adjusted $R^2$ values slows down considerably after the fifth variable.
Figure 5.3 shows the selection results of best subset models of Figure 5.2 based on adjusted $R^2$, Cp and BIC.

We notice that the five variable model is the best of the subset models based on the BIC criteria. However, the improvements in test MSE flatten considerably after the third variable for all “adjusting training error” techniques.

2- Models without clearness index (K) predictor.

Figure 5.4 shows the results of “best subset selection” technique for the $K_d$ response models that do not include K among the predictors.
For the one variable models, opaque cloud cover (OpqC) model is the best. However, in the best two variable model, OpqC leaves and total cloud cover (TotC) enters, relative humidity (RH) enters and continues to the end. In the three variable model, AOD enters and continues to the end, TotC leaves and OpqC enters again. These two variables switch with each other in best subset models, such that one of them is among the predictors. The adjusted $R^2$ values for the models of Figure 5.4, arranged in the same order are below. We notice that cloud cover, RH, and AOD explain 74% of the variability in $K_d$, while the three variable model in Figure 5.2, which includes K, RH and AOD predictors, explains 89% of the variability in $K_d$.

```
reg.summary$adjr2
 [1] 0.6339046 0.7047477 0.7439027 0.7688691 0.7744619 0.7871671 0.7944838 0.7967297
 [9] 0.7978422 0.7984861 0.7983763
```

Figure 5.5 shows the selection results of best subset models of Figure 5.4 based on adjusted $R^2$, Cp and BIC. According to BIC technique, the best model of the subsets is the eight variable model. Meanwhile, for both adjusted $R^2$ and Cp the improvements in test MSE flatten considerably after the eighth variable.

![Graphs showing model selection based on adjusted $R^2$, Cp, and BIC](image)

Figure 5.5: Best $K_d$ response model of the subsets without K predictor, based on adjusted $R^2$, Cp and BIC selection

In this analysis, we excluded H because the correlation matrix shows a negative correlation between H and $K_d$, this is expected since $K_d = \frac{H_d}{H}$. However, we repeated the analysis after adding H to the predictors list. For the models that include predictor K, H enters in the four variable model.
and continues to the end. The improvement in adjusted $R^2$ caused by H is very small (around 0.002) starting from the four variable model. The adjusted $R^2$ values are below.

```r
reg.summary$adjr2
[1] 0.8190461 0.8846756 0.8926206 0.8955941 0.9005483 0.9012674 0.9017730 0.9021283
[9] 0.9030313 0.9031477 0.9031670 0.9031351 0.9030055
```

For the models, which do not include the predictor K, H enters in the three variable model and continues to the end. The improvement in adjusted $R^2$ values caused by H is around 0.02 starting from the three variable model. The adjusted $R^2$ values for these models are below.

```r
reg.summary$adjr2
[1] 0.6339046 0.7047477 0.7602776 0.7915540 0.7933303 0.8002970 0.8125014 0.8169298
[9] 0.8208392 0.8231241 0.8229188 0.8226739
```

Adding H to the predictors list was just for checking and it has no justification. However, H masked the effect of RH, but did not affect AOD and cloud cover predictors. We analyzed the $K_d$ models, excluding H from the predictors list since it does not add significant improvement.

**A- Models with predictor K**

The three variable model includes the predictors K, AOD, and RH. The adjusted $R^2$ value for this model is 0.893. Adding the predictors H$_2$O and H$_0$ will increase adjusted $R^2$ to 0.895. This value is the maximum value we can reach even if we added all the variables to the model. For convenience, we analyzed the three variable model and then compared it with the five variable model.

**Fitting the model**

Figure 5.6 shows the results of fitting the three variable model. The coefficients of the model are very significant for all variables. The most effective predictor is K and it is correlated negatively with $K_d$. This is expected since $K = \frac{H}{H_0}$ and $K_d = \frac{H_d}{H}$. Physically this means that when the proportion of global radiation we receive is large, most of the radiation is direct radiation with a small portion of diffuse radiation. On the other hand, when the proportion of global radiation we receive is small, most of the radiation is diffuse radiation. The second effective predictor is AOD and it is correlated positively with $K_d$. This is normal since diffuse radiation consists of the photons scattered by the atmosphere constituents. RH effect is small and it is correlated negatively with Kd although the correlation matrix showed positive correlation.
Equation 5.1 represents the fitted model:

\[ K_d = 1.036 - 1.220K + 0.735AOD - 0.001RH + \epsilon \]  

\( (5.1) \)

**Diagnostic analysis.**

Figure 5.7 shows four major diagnostic plots for model (5.1). The residuals plot, A, shows no profound pattern of nonlinearity of data or non-constant variance of residuals. Normal Q-Q plot, B, shows no significant deviation from normality, where the majority of the points lie on the Q-Q line. The furthest two point above the Q-Q line represents ND Feb-01 and UT Aug-98; these points are potential outliers. Plot C indicates that there is only one outlier point according to Bonferroni critical value, which equals four, using \( \alpha = 0.05 \). This point belongs to ND Feb-01. The Jacknife residual for UT Aug-98 equals 3.5. Accordingly, it is not considered an outlier. Plot D shows the points with high-leverage. There are no concerns about these points since none of them is outlier.

Finally, Figure 5.8 shows the correlation of error terms \((\epsilon_1, \ldots, \epsilon_n)\) test, using the data of Tallahassee Regional AP –FL (left) and Corpus Christi Intl Arpt –TX (right). No profound correlation exists among the residuals.
Figure 5.7: Diagnostic plots for model (5.1). A- Linearity and constant variance of $\epsilon$ test. B- Normality test ($\epsilon \sim N(0, \sigma^2)$). C- Outliers test. D- High leverage points test.

Figure 5.8: Residuals vs Time for model (5.1). Left is FL station, right is Corpus-TX station
The variance inflation factor (VIF) values of model (5.1) variables are VIF (K) = 3.04, VIF (AOD) = 1.06, and VIF (RH) = 3.12. These values indicate that there is no collinearity problem.

**Comparison with five variable model**

The best five variable model includes K, H₀, RH, AOD, and H₂O (precipitation). Equation 5.2 represents the model.

\[
K_d = 1.008 - 1.215K + 0.700AOD - 0.001RH - 0.008H₂O + 3.004 \times 10^{-6}H₀ + \epsilon \quad (5.2)
\]

All the model coefficients are significant. Precipitation (H₂O) has negative correlation with \( K_d \) as RH does. \( H₀ \) has very small effect and correlates positively with \( K_d \). The VIF values of this model are VIF (K) = 4.21, VIF (H₀) = 3.06, VIF (RH) = 4.45, VIF (AOD) = 1.75, and VIF (H₂O) = 2.44. Since none of the VIF values above five, there is no collinearity problem. However, five variable model does not add significant improvement to the model. Accordingly, the three variable model is preferred due to its simplicity.

**B- Models without predictor K**

We fitted the model with four predictors since the improvement in adjusted \( R^2 \) becomes very small after the four variable model. Notice that after the four variable model, the temperature predictors start to coexist in the models, which causes high collinearity in the model.

**Fitting the model**

The predictors of the best four variable model are OpqC, RH, AOD, and H₀. Figure 5.9 shows the results of fitting the four variable model. The coefficients of the model are very significant for all variables. OpqC, AOD, and RH correlate positively with \( K_d \), while \( H₀ \) has negative but very small correlation with \( K_d \). Excluding K from predictors list increased the positive effect of AOD and unmasked the positive effect of cloud cover. On the other hand, it reversed the signs of \( H₀ \) and RH coefficients. Apart from K, the negative coefficient of \( H₀ \) in this model can be explained as follows: In summer, where \( H₀ \) has large values, the frequency of cloud cover is less, which means less diffuse fraction. The positive coefficient of RH could be due the effect of water vapor molecules in scattering and reflecting the photons. This model explained 77% of \( K_d \) variability.
Equation 5.3 represent the fitted model.

\[ K_d = 0.096 - 9.086 \times 10^{-6} H_0 + 0.035 OpqC + 0.002 RH + 0.894 AOD + \epsilon \quad (5.3) \]

Figure 5.10 shows the diagnostic plots of model (5.3). There are no profound problems recognized. Two outliers exist: WI Jan-01 and ND Jan-99.

Figure 5.10: Diagnostic plots for model (5.3).
Figure 5.11 shows the residuals correlation test. No profound correlation exists among the residuals.

![Correlation of Residuals Test](image)

Figure 5.11: Residuals vs Time for model (5.3)

The VIF values of model (5.3) are $\text{VIF (H}_0\text{)} = 1.85$, $\text{VIF (OpqC)} = 1.89$, $\text{VIF (RH)} = 1.84$, and $\text{VIF (AOD)} = 1.71$. These values indicate that there is no collinearity problem among the variables.

### 5.2.2 Models with Diffuse Radiation ($H_d$) Response

Here we studied the effect of atmospheric parameters on the amount of diffuse radiation received at earth’s surface.

**Variables Selection**

Applying best subset selection technique for $H_d$ response models, we obtained the results shown in Figure 5.12. The first variable to enter is $H_0$ and it continues to the end. For two variable model, $H$ enters and continues to the end. The third variable to enter is AOD and it continues to the end. The fourth variable to enter is TotC and it continues to the end. The fifth variable to enter is OpqC and it continues to the end. However, due to the high collinearity between TotC and OpqC, the improvement in adjusted $R^2$ of the five variable model is expected to be very small.
Below is the adjusted $R^2$ values for the models in Figure 5.12, arranged in the same order.

We notice that the improvement in adjusted $R^2$ values becomes very small after the four variable model. The reason is the collinearity between the cloud cover variables, and among temperature variables.

```
reg.summary$adjr2  
[1] 0.7704451 0.8880125 0.9198552 0.9230609 0.9249377 0.9257798 0.9267174 0.9276319
[9] 0.9281422 0.9287802 0.9287549 0.9286752
```

Figure 5.13 shows the best model of the subsets according to adjusted $R^2$, $Cp$, and BIC. All the techniques selected ten variable model as the best model. However, we notice very small improvements in the models after the fourth variable in all techniques.

Figure 5.12: Best subset selection results for the models with $H_d$ response

Figure 5.13: Best model of $H_d$ response subsets based on adjusted $R^2$, $Cp$ and BIC selection
Four variable model

Based on the above analysis, we fitted the four variable model. The predictors for the best four variable model are $H_0$, $H$, AOD, and TotC.

**Fitting the model**

Figure 5.14 shows the results of fitting the four variable model of $H_d$ response. All the coefficients are very significant. The most effective variable is AOD then TotC. $H_d$ has positive correlation with all predictors except $H$. This indicates that as the global radiation we receive increases the diffuse part decreases, where most of the global radiation becomes direct radiation. However, the coefficients of $H$ and $H_0$ predictors are very small compared to AOD and TotC.

![Call: lm(formula = H_d ~ H0 + H + AOD + TotC, data = MDAllVar)]

![Residuals:
Min 1Q Median 3Q Max
-432.80 -93.41 -2.88 78.40 843.69]

![Coefficients:
(Intercept) -501.7518 34.5280 14.532 < 2e-16 ***
H0 0.3073 0.0100 30.643 < 2e-16 ***
H -0.1919 0.0150 -12.788 < 2e-16 ***
AOD 3291.1998 194.6781 16.906 < 2e-16 ***
TotC 32.7888 5.9049 5.553 3.97e-08 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 166.7 on 715 degrees of freedom
Multiple R-squared: 0.9235, Adjusted R-squared: 0.9231
F-statistic: 2158 on 4 and 715 DF, p-value: < 2.2e-16]

Figure 5.14: The results of fitting $H_d$ against $H_0$, $H$, AOD, and TotC.

Equation 5.4 represents the fitted model.

$$H_d = -501.752 + 0.307 \ H_0 - 0.192 \ H + 3291.200 \ AOD + 32.789 \ TotC + \epsilon \quad (5.4)$$

**Diagnostic analysis.**

Figure 5.15 shows the diagnostic plots of model (5.4). Plot (A) shows no profound pattern of nonlinearity of data or non-constant variance of residuals. Plot (B) shows some deviation from
normality in the upper right corner of the plot. These points are potential outliers. The furthest point represents UT Jun 2002. Plot (C) shows several outliers according to Bonferroni critical value, which equals four, using $\alpha = 0.05$. These outliers belong to UT Aug-98, UT Jun-99, UT Jun-00, UT Jun-01, UT Jun-02, and UT Aug-02. UT Jul-00 has Studentized residual equals 3.9, so it is almost outlier. However, these outliers are not problem, since none of them has high-leverage and we have 720 points. Plot (D) shows no concerns regarding high-leverage points.

![Residuals vs Fitted](image1)

![Normal Q-Q Plot](image2)

![Jacknife Residuals vs Fitted](image3)

![Jacknife Residuals vs Leverage](image4)

Figure 5.15: Diagnostic plots for model (5.14). A- Linearity and constant variance of $\epsilon$ test. B- Normality test ($\epsilon \sim N(0, \sigma^2)$). C- Outliers test. D- High leverage points test.

Figure 5.16 shows the correlation of residuals plot for model (5.4). There is no profound correlation among the residuals.

The VIF values of model (5.4) are VIF (H$_0$) = 17.75, VIF (H) = 20.38, VIF (AOD) = 1.70, and VIF (TotC) = 2.69. There is collinearity problem related to H and H$_0$ variables. Consequently, we reanalyzed the model and explored what effect excluding H or H$_0$ would have.
Comparison with other models

1- Model without predictor $H_0$

Figure 5.17 shows the results of best subset selection technique for $H_d$ response models that exclude the $H_0$ predictor, along with the adjusted $R^2$ values of these models.

```
1 subsets of each size up to 11
Selection Algorithm: exhaustive

1  (1)                                 1
2  (1)                                 2
3  (1)                                 3
4  (1)                                 4
5  (1)                                 5
6  (1)                                 6
7  (1)                                 7
8  (1)                                 8
9  (1)                                 9
10 (1)                                 10
11 (1)                                 11
```

> reg.summary$adjr2

|   |      |      |      |      |      |      |      |      |      |      |
|---|------|------|------|------|------|------|------|------|------|
|   | 0.5428241 | 0.7462009 | 0.8222666 | 0.8662779 | 0.8711739 | 0.8714296 | 0.8714715 | 0.8718270 | 0.8718270 |
| [9] | 0.8718809 | 0.8717484 | 0.8715799 |

Figure 5.17: Best subset selection results for $H_d$ response models that exclude predictor $H_0$
The first variable to enter is AOD. In the two variable model, AOD leaves and RH enters, H enters and continues to the end. In the three variable model, RH leaves, AOD re-enters and continues to the end, and TotC enters and continues to the end. In the four variable model, RH re-enters and continues to the end. The improvement in adjusted $R^2$ values becomes very small after the four variable model.

Figure 5.18 shows the results of fitting the four variable model. All the variables are very significant and correlated positively with $H_d$. The least effective parameter is H. This model explains 86.6% of the variability in $H_d$, while model (5.4) explains 92.3% of $H_d$ variability.

Equation 5.5 represent the fitted model.

$$H_d = -127.7 \times 10 + 0.3 \ H + 419.5 \times 10 \ AOD + 103.4 \ TotC + 10.6 \ RH + \epsilon \quad (5.5)$$

2- Model without predictor H

Figure 5.19 shows the results of best subset selection technique for $H_d$ response models, which exclude the H predictor, along with the adjusted $R^2$ values of these models. The first variable to enter is $H_0$ and it continues to the end. In the two variable model, TotC enters and continues to the end. In the three variable model, AOD enters and continues to the end. In the four variable model, RH enters and continues to the end. The improvement in adjusted $R^2$ values becomes very small after the four variable model.
According to adjusted $R^2$ values, we notice that the performance of models that exclude predictor $H$ is better than those exclude $H_0$. This is expected since the collinearity of $H_0$ is less than the collinearity of $H$.

Figure 5.20 shows the results of fitting the four variable model of Figure 5.19. All the variables are very significant and correlated positively with $H_d$. The least effective parameter is $H_0$. This model explains 91.4 % of the variability in $H_d$ compared to 86.6 % for model (5.5) that excludes $H_0$, and 92.3 % for model (5.4) that includes both $H$ and $H_0$. 

Figure 5.20: The results of fitting $H_d$ against $H_0$, AOD, TotC, and RH.
Equation 5.6 represent the four variables model

\[ H_d = -871.8 + 0.2 H_0 + 338.2 \times 10 AOD + 64.2 TotC + 4.4 RH + \epsilon \quad (5.6) \]

Although model (5.4) has better performance than model (5.6), model (5.6) has the advantage that all its predictors are easy to measure or calculate. We tested the prediction ability of both models.

Finally, we analyzed models with response \( \log (H_d) \) to see if these models enhance the ability of predicting \( H_d \).

### 5.2.3 Models with \( \log (H_d) \) Response

We noticed that model (5.1) explained only 89.3% of the variability of diffuse fraction (\( K_d \)). To enhance the model performance, we used \( \log (K_d) \) in the same manner we did with \( \log (K) \) in section 4.2.2. However, this time we have both \( K_d \) and \( K \) in model (5.1). From the analysis, we know that \( K_d \propto -K \). We proceed as following:

\[
\begin{align*}
K_d & \propto -K \\
& \rightarrow \log(K_d) \propto -\log(K) \\
& \rightarrow \log(H_d) - \log(H) \propto -\log(H) + \log(H_0) \\
& \rightarrow \log(H_d) \propto \log(H_0)
\end{align*}
\]

(5.7)

Based on Equation 5.7, the response is \( \log(H_d) \). \( \log(H_0) \) is among the predictors, while \( \log(H) \) is not.

**Variables Selection**

Figure 5.21 shows the results of best subset selection technique for \( \log (H_d) \) response models, along with the adjusted \( R^2 \) values of these models. The selection of best subset technique for the models with one, two, three, and four variables is the same as in Figure 5.19 (\( H_d \) response models that exclude predictor \( H \)). However, after the four variable model, the selection differs from those in Figure 5.19. We notice that there is enhancement in adjusted \( R^2 \) values due to using \( \log (H_d) \) response. As before, the improvement in adjusted \( R^2 \) values becomes very small after the four variable model. Accordingly, we fitted the four variable model.
Fitting the model

Figure 5.22 shows the results of fitting the four variable model. The coefficients of all the predictors are very significant. All the predictors correlate positively with log (H_d). The model explains 93.43% of the variability of log (K_d).

Equation 5.8 represents the model.

\[
\text{log} (H_d) = -2.234 + 0.993 \text{log}(H_0) + 2.232 \text{AOD} + 0.045 \text{TotC} + 0.003 \text{RH} + \epsilon \quad (5.8)
\]
Diagnostic analysis.

Figure 5.23 shows the diagnostic plots for model (5.8). No profound problems exist that affect the validity of the model. However, we have one outlier point that belongs to UT Aug 2002.

![Diagnostic plots for model (5.8)](image)

Figure 5.23: Diagnostic plots for model (5.8)

Figure 5.24 shows the correlation of residuals plot for model (5.8). There is no profound correlation among the residuals.

![Correlation of Residuals Test](image)

Figure 5.24: Residuals vs Time for model (5.8). Left side is FL station Data, Right side is Corpus-TX station
Finally, the VIF values of model (5.8) variables are VIF (log (H₀)) = 1.70, VIF (TotC) = 1.80, VIF (RH) = 1.74, and VIF (AOD) = 1.56. These values indicate that there is no collinearity problem among the variables.

5.3 Prediction Analysis

We tested the prediction ability of three models fitted in the previous section. These models are:

\[ H_d = -501.752 + 0.307 \ H_0 - 0.192 \ H + 3291.200 \ AOD + 32.789 \ TotC + \epsilon \] (5.4)

\[ H_d = -871.8 + 0.2 \ H_0 + 338.2 \times 10 \ AOD + 64.2 \ TotC + 4.4 \ RH + \epsilon \] (5.6)

\[ \text{log} (H_d) = -2.234 + 0.993 \log(H_0) + 2.232 \ AOD + 0.045 \ TotC + 0.003 \ RH + \epsilon \] (5.8)

The accuracy and performance of the models were assessed using the following statistical indicators, defined in section 4.3:

1- The adjusted coefficient of determination (adjusted R²)
2- The mean percentage error (MPE)
3- The mean absolute percentage error (MAPE)
4- The root mean squared error (RMSE)
5- The mean bias error (MBE)

The test data is the same data used to test the performance of global radiation models; see Table 4.1 in Chapter 4. The validation results of models (5.4), (5.6), and (5.8) are shown in Table 5.1.

Table 5.1: The accuracy of diffuse models based on the statistical error tests.

<table>
<thead>
<tr>
<th>Model</th>
<th>MPE</th>
<th>MAPE</th>
<th>RMSE</th>
<th>MBE</th>
<th>Adjusted R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>(5.4) H_d response</td>
<td>1.921%</td>
<td>7.772%</td>
<td>135.174</td>
<td>32.131</td>
<td>0.944</td>
</tr>
<tr>
<td>(5.6) H_d response</td>
<td>2.003%</td>
<td>9.069%</td>
<td>152.922</td>
<td>28.669</td>
<td>0.928</td>
</tr>
<tr>
<td>(5.8) log (H_d) response</td>
<td>1.508%</td>
<td>8.061%</td>
<td>153.420</td>
<td>16.211</td>
<td>0.928</td>
</tr>
</tbody>
</table>

As mentioned before, the major statistical indicator is RMSE (the square root of test MSE). Table 5.1 shows that the lowest RMSE is for model (5.4), which has the predictors H₀, H, AOD, and TotC. Accordingly, this model has the best prediction ability. Models (5.6) and (5.8) have
similar prediction performance, since their values of RMSE are comparable. These two models have the same predictors, which are H₀, AOD, TotC, and RH. However, model (5.6) is preferred since mathematically it is simpler.

Adjusted R² values confirm the above results. Model (5.4) has the highest adjusted R² value, that is, the best performance, and models (5.6) and (5.8) have the same adjusted R² value, which means similar performance. Interestingly, for model (5.4) the training adjusted R² is 0.923, while it is 0.944 for the test data. On the other hand, the training adjusted R² for model (5.6) is 0.914 and for model (5.8) is 0.934, while they have the same test adjusted R² value, which is 0.928. In general, the values of adjusted R² indicate that diffuse models perform good, but not as good as global radiation models. Diffuse models were able to explain around 93% of the variability of diffuse radiation.

As mentioned before, models (5.6) and (5.8) have similar predictors, but model (5.6) is simpler mathematically, since it deals with H_d and H_0 instead of their logarithms. Since these models have similar performance, based on RMSE and adjusted R² indicators, it is clear that using natural logarithm of H_d and H_0 did not enhance the performance of the model. Actually, RMSE for model (5.6) is slightly smaller than RMSE for model (5.8).

All the models have positive MBE values. This indicates that all the models tend to overestimate the diffuse radiation. All the models have small positive MPE values. This indicates that the percentage error of overestimated values slightly prevails the percentage error of underestimated values. As mentioned before, a drawback of MPE (as well MBE) that an overestimation of a response can be hidden by an underestimation of another response. Consequently, for a realistic estimation of the average percentage deviation between predicted and measured values of diffuse radiation, we look at MAPE values. Table 5.1 shows that the MAPE values of the diffuse models are between 8 % and 9 %. These values are considered acceptable although they are larger than the MAPE values of global radiation models.

In the same manner as in section 4.3, the comparison among the models performance was based on the predicted and measured values of H_d. This includes the log (H_d) response model, where the predicted values of log (H_d) were converted to H_d values, before calculating the statistical error tests.
5.4 Discussion and Conclusions

Using MLR to understand the interaction between diffuse radiation and atmospheric parameters we found the following:

1- The main predictor that affect the value of diffuse fraction (K_d) is clearness index (K). This predictor explains 82% of the K_d variability. K_d correlates negatively with K. This indicates that the greater the percentage of radiation we receive at earth’s surface, the smaller the diffuse proportion is. That is, for large values of K, most of the global radiation is direct with small fraction of diffuse radiation.

2- In the presence of predictor K, the main atmospheric factors that affect K_d are aerosol optical depth (AOD) and relative humidity (RH). In the absence of predictor K, the main atmospheric factors that affect K_d are cloud cover, RH, and AOD. This indicates that K masks the effect of cloud cover.

3- K_d has positive correlation with AOD and cloud cover. This can be attributed to the role of aerosols and clouds in scattering the photons (diffuse radiation results from scattered photons). K_d has negative correlation with RH in the presence of predictor K and positive correlation with RH in the absence of predictor K. Accordingly, predictor K disrupted the effect of RH on diffuse fraction and made it vague. However, the positive correlation between K_d and RH is more likely since there is positive correlation between diffuse radiation (H_d) and RH.

4- The main atmospheric parameters that affect the amount of diffuse radiation received at earth’s surface (H_d) are total cloud cover (TotC), AOD, and RH. TotC and AOD have positive correlation with H_d and large coefficients compared to other predictors. RH has positive correlation with H_d. This could be explained in terms of the scattering effect of water vapor molecules or water droplets in the atmosphere on photons.

5- H_d has positive correlation with the extraterrestrial radiation (H_0); this is natural since H_0 represent the radiation source. H_d has negative correlation with the global radiation (H). This indicates that the larger the amount of global radiation we receive at earth’s surface, the smaller the amount of diffuse radiation we receive, where most of the radiation will be direct radiation. The negative correlation between K_d and K, and between H_d and H can be explained as follows: since we receive large amounts of global radiation in clear sky
days, the absence of clouds reduces the photons scattering and consequently the diffuse radiation.

6- The importance of atmospheric parameter effect on diffuse radiation depends on the response variable of the model and the accompanied predictors. Table 5.2 shows the effective parameters, in descending importance, for the major models.

Table 5.2: The most important parameters in descending order for the major diffuse models.

<table>
<thead>
<tr>
<th>Order</th>
<th>The response of the model</th>
<th>K_d</th>
<th>Log (H_d)</th>
<th>H_d</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Predictor K In No predictor K</td>
<td>K_d</td>
<td>Log (H_d)</td>
<td>H_d</td>
</tr>
<tr>
<td>First</td>
<td>K Cloud cover Log (H_0)</td>
<td>H_0</td>
<td>H_0</td>
<td></td>
</tr>
<tr>
<td>Second</td>
<td>AOD RH TotC H</td>
<td>TotC</td>
<td>AOD</td>
<td></td>
</tr>
<tr>
<td>Third</td>
<td>RH AOD AOD TotC</td>
<td>AOD</td>
<td>RH</td>
<td></td>
</tr>
<tr>
<td>Fourth</td>
<td>OpqC H_0 RH AOD</td>
<td>RH</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We notice that, for proportion response (K_d), cloud cover occupies the first place in the absence of predictor K, and the fourth place in the presence of predictor K. RH is also affected by K, but in a lighter manner, where it moves from the third position to the second in the absence of K. This implies that, the predictor K disturbs or wraps the effect of cloud cover and relative humidity. For the H_d and log (H_d) responses, H_0 occupies the first place. This is expected since H_0 represents the source. In the absence of predictor H, TotC occupies the second place, AOD the third, and RH the forth. This order is reasonable, since it is commensurate with the contribution of these parameters to the photons scattering. In the presence of H predictor, H takes the second place and RH effect is totally wrapped or disrupted.

7- Adding wind speed (WS), water precipitation (H_2O), and temperature variables to the model predictors did not improve the model performance significantly. This indicates that their effect on diffuse radiation is negligible.

8- After the forth predictor, adding more variables did not improve the models performance significantly. This is attributed mainly to the collinearity problem among the temperature variables and between total and opaque cloud covers.
9- The validation test showed that all diffuse models have good prediction ability with model (5.4) has the best performance. However, this model includes H among its predictors, which is costly to measure. Excluding H from predictors list lowered the ability of prediction slightly. Consequently, model (5.6), which has the predictors $H_0$, TotC, AOD, and RH, is a good and convenient option. All its predictors are easy to measure or calculate and it explains 93% of the variability in $H_d$ according to the test data. All diffuse models overestimate diffuse radiation.

Our work is similar to Li et al. (2012) work. The authors developed two models for estimating the monthly average daily diffuse solar radiation in China. They used $T$, RH, $K$ and sunshine fraction variables to fit the models. They concluded that incorporating ambient temperature and relative humidity to the models with $K$ and sunshine duration predictors could generally improve the models estimates. In our analysis we concentrated on identifying the atmospheric parameters that physically and substantially affect the amount of diffuse radiation. The objective was to fit simple models that able to capture the effect of climate variability on diffuse radiation and predict it effectively. The average RMSE for our diffuse models equal to 147.17 Watthour/m², which is equivalent to 0.53 MJ/m². The average RMSE for Li et al. models equal to 0.66 MJ/m². Accordingly, the performance of our diffuse models is better or at least comparable to the performance of Li et al. models.
6.1 Regularized Regression

In Chapters 4 and 5, we used “Best Subset Selection” technique to determine the atmospheric variables that most affect the amount of global and diffuse radiation received at earth’s surface. Alternately, we can fit a model containing all the predictors using a technique that constrains or regularizes the coefficient estimates by shrinking the coefficient estimates towards zero. The two best-known techniques for shrinking the regression coefficients towards zero are “ridge regression” and “the lasso” (James et al. 2014).

\textit{Ridge regression and the lasso}

Ridge regression is very similar to least squares, except that there is a shrinkage penalty term added to the residual sum of squares (RSS). The coefficients are estimated by minimizing the quantity in the following equation:

\[
\sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2 \quad (6.1)
\]

where \( y_i \) is the \( i \)th response, \( \beta_0 \) is the intercept coefficient of the model, \( \beta_j \) is the coefficient of predictor or variable \( j \), \( x_{ij} \) is the \( i \)th observation for the variable \( j \), and \( \lambda \geq 0 \) is a tuning parameter, to be determined separately.

Equation 6.1 trades off two different criteria. The first, represented by RSS term, is the regular least squares criterion that seeks coefficient estimates, which fit the data well, by making the RSS small. The second criterion is represented by the second term \( \lambda \sum_{j=1}^{p} \beta_j^2 \), which is called a shrinkage penalty. \( \lambda \sum_{j=1}^{p} \beta_j^2 \) is small when \( \beta_1, \ldots, \beta_p \) are close to zero. The relative impact of these two terms on the regression coefficient estimates is controlled by the tuning parameter \( \lambda \). When \( \lambda = 0 \), the penalty term has no effect, and ridge regression produces the least squares estimates. However, as \( \lambda \to \infty \), the impact of the shrinkage penalty grows, and the ridge regression coefficient estimates approach zero. However, ridge regression will always generate a model that contains all the variables. It tends to reduce the magnitudes of the coefficients, but does not result in exclusion of any of the variables. The reason is that ridge regression was designed originally to
serve chemical engineering analysis, where the variables have collinearity among them, but all of them are important and should be included in the analysis. Ridge regression is able to reduce the collinearity among the variables without excluding any of them, by its shrinkage penalty. However, reducing the magnitude of coefficients without excluding any of the variables is disadvantage in the cases where variable selection is required. An alternative to ridge regression that overcomes this disadvantage is the lasso.

The lasso estimates the coefficients of a model by minimizing the following quantity:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| = RSS + \lambda \sum_{j=1}^{p} |\beta_j|  \quad 6.2$$

Comparing Equation 6.2 with Equation 6.1, we notice that the lasso and ridge regression have similar formulation. The only difference is that the $\beta_j^2$ term in the ridge regression penalty is replaced by $|\beta_j|$ in the lasso penalty. The lasso shrinks the coefficient estimates towards zero, as ridge regression does. However, the penalty in the lasso has the effect of forcing some coefficients to be exactly zero, when $\lambda$ is sufficiently large. Hence, the lasso performs variable selection like best subset selection technique (James et al. 2014).

**Global and diffuse models using ridge regression and the lasso**

We applied ridge regression and the lasso on our data to build models for global radiation and diffuse radiation. The models were fitted, tested and compared with several MLR models. Table 6.1 shows the results of global radiation models. The lasso, with optimized $\lambda = 3.4$, selected seven variables. While ridge regression, with optimized $\lambda = 192.0$, did not exclude any variable, as expected. The variables selected by the lasso are similar to the variables selected by the seven variable MLR model, except for one variable, where the lasso selected aerosol optical depth (AOD) while the seven variable MLR model selected TDaylight. To compare the models performance, we used the same test data of Chapter 4, which is illustrated in Table 4.1. Table 6.1 shows the RMSE and adjusted R2 values for five MLR models, the lasso and ridge regression. We notice that the performance of seven variable MLR model is slightly better than the lasso. The worst performance is for ridge regression, while the best performance is for the five variable MLR model. As mentioned before, the performance of MLR models is comparable, starting from the three variable model. For the simplicity, the three variable model is preferred, where we need only the
data of total cloud cover and relative humidity, beside the calculated $H_0$. Table 6.2 shows the results of diffuse radiation models. Both the lasso, with optimized $\lambda = 0.06$, and ridge regression, with optimized $\lambda = 57.86$, did not exclude any variable. The best performance is for the lasso, which has twelve variables. However, its performance is comparable to the four and five variable MLR models, consequently they are preferred for their simplicity. The worst performance is for ridge regression. Accordingly, “Best Subset Selection” technique is more effective and efficient in choosing the variables that most affect the amount of global and diffuse radiations.

6.2 Autoregressive Analysis for the Residuals

In Chapters 4 and 5, we tested the existence of correlation among the residuals of the fitted models. There was no profound correlation among the residuals for all fitted models. In this section, we investigated the subject further by examining the existence of autoregressive (AR) structure in the residuals. We examined the existence of AR structure in the residuals of two global radiation models, namely, model (4.12) and model (4.13):

$$\log(H) = -0.320 + \log(H_0) - 0.049OpqC - 0.004RH + .012T_{Range} + \epsilon$$  \hspace{1cm} (4.12)

$$H = 2.008 \times 10^3 + 0.583H_0 - 1.522 \times 10^2TotC - 24.09RH + 8.035T + \epsilon$$  \hspace{1cm} (4.13)

Figure 6.1 shows the plots of autocorrelation function, which gives the correlation of a time series with its own lagged values, for the residuals of model (4.12). The x-axis of the plot represents the lagged time and the y-axis represents autocorrelation function values (the correlation between the residuals of the first month and the residuals of the lagged months). The plots are for the stations: Tallahassee Regional-FL, Corpus Christi Intl Arpt-TX, Madison Dane Co Regional Arpt-WI, and Eugene Mahlon Sweet Arpt-OR. Figure 6.1 shows no obvious AR structure in the residuals of the stations. However, it reveals a periodic trend in the data, similar to sine function. This trend is related to the seasons of the year. We notice a negative correlation in the middle of the year (5, 6, and 7 lags) and a positive correlation at the end of the year and at the beginning of the next year (Lags 10, 11, and 12). This trend is not clear in FL station because of its humid subtropical climate.
Table 6.1: The coefficients and performance of MLR models, the lasso ($\lambda=3.4$) and ridge regression ($\lambda=192.0$) for global radiation.

<table>
<thead>
<tr>
<th>Regression Selection Method</th>
<th>H</th>
<th>TotC</th>
<th>OpqC</th>
<th>H2O</th>
<th>AOD</th>
<th>T_{Min}</th>
<th>T</th>
<th>T_{Daylight}</th>
<th>T_{Range}</th>
<th>RH</th>
<th>WS</th>
<th>No. of var.</th>
<th>RMSE</th>
<th>Adj. $R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLR Best subset selection</td>
<td>0.61</td>
<td>-163.7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-23.47</td>
<td>0</td>
<td>3</td>
<td>310.39</td>
<td>0.976</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ridge Lasso</td>
<td>0.57</td>
<td>-149.21</td>
<td>0</td>
<td>-123.44</td>
<td>247.21</td>
<td>0</td>
<td>20.65</td>
<td>0</td>
<td>12.87</td>
<td>-19.31</td>
<td>7</td>
<td>306.38</td>
<td>0.976</td>
<td></td>
</tr>
<tr>
<td>Ridge Ridge</td>
<td>0.41</td>
<td>-62.11</td>
<td>-52.23</td>
<td>-116.25</td>
<td>3517</td>
<td>15.11</td>
<td>19.97</td>
<td>13.98</td>
<td>58.18</td>
<td>-19.31</td>
<td>11</td>
<td>357.02</td>
<td>0.967</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.2: The coefficients and performance of MLR models, the lasso ($\lambda=0.06$) and ridge regression ($\lambda=57.86$) for diffuse radiation.

<table>
<thead>
<tr>
<th>Regression Selection Method</th>
<th>H</th>
<th>TotC</th>
<th>OpqC</th>
<th>H2O</th>
<th>AOD</th>
<th>T_{Min}</th>
<th>T</th>
<th>T_{Daylight}</th>
<th>T_{Range}</th>
<th>RH</th>
<th>WS</th>
<th>No. of var.</th>
<th>RMSE</th>
<th>Adj. $R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLR Best subset selection</td>
<td>0.31</td>
<td>-0.19</td>
<td>32.79</td>
<td>0</td>
<td>0</td>
<td>3291.2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>135.17</td>
<td>0.944</td>
<td></td>
</tr>
<tr>
<td>Ridge Ridge</td>
<td>0.32</td>
<td>-0.18</td>
<td>190.84</td>
<td>-165.51</td>
<td>-54.32</td>
<td>3298.6</td>
<td>-3.1</td>
<td>-70.1</td>
<td>74.3</td>
<td>-16.8</td>
<td>5.01</td>
<td>12</td>
<td>133.93</td>
<td>0.945</td>
</tr>
<tr>
<td>Ridge Ridge</td>
<td>0.12</td>
<td>0.06</td>
<td>51.02</td>
<td>23.56</td>
<td>9.39</td>
<td>3857.2</td>
<td>1.1</td>
<td>1.3</td>
<td>1.1</td>
<td>3.9</td>
<td>9.86</td>
<td>12</td>
<td>166.12</td>
<td>0.914</td>
</tr>
</tbody>
</table>
Figure 6.2 shows the plots of autocorrelation function for the residuals of model (4.13) for the same stations above. Figure 6.2 shows no obvious AR structure in the residuals of model (4.13). However, it reveals the periodic trend related to the seasons of the year. As in Figure 6.1, this periodic trend is not clear in FL station plot because of its humid subtropical climate.
Since the residuals of MLR models we discussed do not have a time series structure, the assumption of independent errors made in ordinary least squares regression is not violated and there is no need to adjust estimated regression coefficients and standard errors for AR structure in the errors.

### 6.3 Leave Out One Station Cross Validation

In Chapter 4 and 5, we tested the performance of global and diffuse models using the data of five stations that were part of the stations included in fitting the models. In this section, we fitted and validated the performance of two models using Leave Out One Station Cross Validation (LOOSCV). The first is the global radiation model, represented by Equation 6.3 and the second is the diffuse radiation model, represented by Equation 6.4.

\[
H = \beta_0 + \beta_1 H_0 + \beta_2 TotC + \beta_3 RH + \beta_4 T + \epsilon \quad (6.3)
\]

\[
H_d = \beta_0 + \beta_1 H_0 + \beta_2 TotC + \beta_3 RH + \beta_4 AOD + \epsilon \quad (6.4)
\]

In LOOSCV method, we fitted the models using the data of all stations excluding one station, then we tested the performance of the model using the data of the excluded station. We repeated this process 11 additional times, excluding a different station each time. The objective of this method is to inspect the variation in model performance from station to station. Table 6.3 shows the results for global radiation model (6.3)

Table 6.3: LOOSCV for model (6.3); station name represents the station used in testing the model.

<table>
<thead>
<tr>
<th>Stat.</th>
<th>FL</th>
<th>TX.Co</th>
<th>TX.Au</th>
<th>TX.Ab</th>
<th>TX.Am</th>
<th>NM</th>
<th>NV</th>
<th>VA</th>
<th>UT</th>
<th>WI</th>
<th>OR</th>
<th>ND</th>
<th>Ave.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train. Adjr²</td>
<td>0.978</td>
<td>0.977</td>
<td>0.977</td>
<td>0.976</td>
<td>0.976</td>
<td>0.976</td>
<td>0.976</td>
<td>0.975</td>
<td>0.976</td>
<td>0.975</td>
<td>0.977</td>
<td>0.976</td>
<td>0.976</td>
</tr>
<tr>
<td>Test RMSE</td>
<td>355.6</td>
<td>301.6</td>
<td>314.7</td>
<td>247.5</td>
<td>262.4</td>
<td>315.5</td>
<td>374.8</td>
<td>265.0</td>
<td>261.3</td>
<td>297.2</td>
<td>418.1</td>
<td>264.0</td>
<td>306.5</td>
</tr>
<tr>
<td>Test Adjr²</td>
<td>0.919</td>
<td>0.952</td>
<td>0.955</td>
<td>0.975</td>
<td>0.976</td>
<td>0.964</td>
<td>0.963</td>
<td>0.966</td>
<td>0.983</td>
<td>0.970</td>
<td>0.958</td>
<td>0.981</td>
<td>0.964</td>
</tr>
</tbody>
</table>
We notice from Table 6.3 that the performance of model (6.3) varies slightly from station to station. FL station has the lowest adjusted R² value at 0.92. However, its RMSE value is lower than those of NV and OR stations, which have adjusted R² values around 0.96. As mentioned before, Florida is distinguished by its humid subtropical climates.

Table 6.4 shows the analysis for diffuse model (6.4). The variation from one station to another is obvious. Stations NV and UT have very low adjusted R² values compared to other stations. This can be attributed to the dry climates which distinguish these two stations from other stations. NM station has adjusted R² value equal to 0.8, while the other stations have adjusted R² values above 0.9. NM has arid to semiarid climates. Further analysis needs to be done to explore the variation in models performance from station to station.

Table 6.2: LOOSCV for model (6.4); station name represents the station used in testing the model.

<table>
<thead>
<tr>
<th>Stat.</th>
<th>FL</th>
<th>TX.Co</th>
<th>TX.Au</th>
<th>TX.Ab</th>
<th>TX.Am</th>
<th>NM</th>
<th>NV</th>
<th>VA</th>
<th>UT</th>
<th>WI</th>
<th>OR</th>
<th>ND</th>
<th>Ave.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Train. Adjr²</td>
<td>0.910</td>
<td>0.910</td>
<td>0.914</td>
<td>0.913</td>
<td>0.912</td>
<td>0.916</td>
<td>0.918</td>
<td>0.905</td>
<td>0.947</td>
<td>0.911</td>
<td>0.914</td>
<td>0.911</td>
<td>0.915</td>
</tr>
<tr>
<td>Test RMSE</td>
<td>119.6</td>
<td>130.9</td>
<td>162.2</td>
<td>130.5</td>
<td>117.9</td>
<td>215.5</td>
<td>233.5</td>
<td>85.5</td>
<td>438.7</td>
<td>136.7</td>
<td>176.3</td>
<td>144.5</td>
<td>174.3</td>
</tr>
<tr>
<td>Test Adjr²</td>
<td>0.954</td>
<td>0.938</td>
<td>0.898</td>
<td>0.975</td>
<td>0.942</td>
<td>0.790</td>
<td>0.513</td>
<td>0.984</td>
<td>0.582</td>
<td>0.947</td>
<td>0.907</td>
<td>0.939</td>
<td>0.864</td>
</tr>
</tbody>
</table>
CHAPTER 7: GENERAL CONCLUSIONS AND FUTURE WORK

7.1 General Conclusions

We applied MLR analysis on part of measured data of National Solar Radiation Database 1991-2010 Update, to understand the coupling of global and diffuse radiation with climatic variability. From the results of analyses, we concluded the following:

1- The major atmospheric parameters that affect the amount of global radiation at earth’s surface are cloud cover, relative humidity, and average temperature. For the diffuse radiation, the major atmospheric parameters are cloud cover, aerosol optical depth, and relative humidity.

2- Global radiation has negative correlation with cloud cover and relative humidity, and positive correlation with temperature. On the other hand, diffuse radiation has positive correlation with cloud cover, aerosol optical depth, and relative humidity. These results can be explained based on the effect of cloud cover, aerosols, and water vapor molecules in the atmosphere. These constituents reflect, absorb and scatter the photons of radiation. The reflection and absorption of photons dissipate part of the solar radiation, which reduces the global radiation amount. The scattering of photons results in diffuse radiation.

3- Wind speed and precipitation have insignificant effect on global and diffuse radiation. Temperature variables have insignificant effect on diffuse radiation.

4- Based on the statistical error tests, calculated using the test data, MLR models of global radiation have better prediction performance than the MLR models of diffuse radiation. Accordingly:
   a. Linear models are excellent approximation for the relationship between global radiation and atmospheric parameters. A linear model with the predictors cloud cover, relative humidity and extraterrestrial radiation \( H_0 \) is able to account for 98% of global radiation variability.
   b. Linear models are a very good approximation for the relationship between diffuse radiation and atmospheric parameters. A linear model with the predictors total cloud cover, aerosol optical depth, relative humidity, and \( H_0 \) is able to account for around 93% of diffuse radiation variability.
c. Using nonlinear terms or nonlinear models might enhance the performance of diffuse radiation.

7.2 Technical Contributions

Previous research focused on fitting models to predict the global and diffuse radiation. Most of the models have two inherent limitations: Generalization over spatial and temporal scale, and high uncertainty given there are limited accounts of inclusion of climatic variables. In this research, our focus was to determine A and B factors of the atmosphere that most affect the amount of global and diffuse radiation received at earth’s surface, using the data of different regions and climates in the USA. Based on the results, we fitted models using these effective atmospheric parameters to estimate global and diffuse radiation. These models can be used to estimate global radiation in places, where solar radiation data is not available. The results showed that global radiation could be estimated efficiently using only two atmospheric parameters, namely, the cloud cover and relative humidity.


Based on the data of average cloud cover and relative humidity in any region, we can estimate the average amount of solar radiation received in this region. Consequently, we can determine if implementing a specific kind of solar radiation harvest technology such as photovoltaic systems and concentrated solar power in this region is practical and economical.

For example, a certain photovoltaic system works efficiently in the range from X to Y of global radiation. From the fitted models, we can build a chart that shows the range of values for cloud cover and relative humidity, which result in global radiation from X to Y. This chart can be used to give an indication if solar radiation technology can be used in a certain place by comparing its cloud cover and relative humidity data with chart data.

7.4 Future Work

For further understanding of the coupling of global and diffuse radiation with atmospheric parameter, we recommend the following:
1- Expand the data to include wide variety of climate conditions, such as tropical and subarctic climates.

2- Redo the analysis using hourly data to investigate the effect of averaging on the detailed relationships.

3- Conduct different methods of statistical learning, such as Multivariate Multiple Regression, where the response is a linear combination of global and diffuse radiation and the predictors are linear combination of the atmospheric parameters.

4- Enhance the prediction ability of diffuse radiation models by using nonlinear terms or nonlinear models such as Regression Splines or Generalized Additive Models. The focus here is on prediction not interpretation.
APPENDIX

Stations and Quality of data used in the research.

<table>
<thead>
<tr>
<th>Station Code</th>
<th>Location Details</th>
<th>Data Quality</th>
<th>Year Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>722140</td>
<td>TALLAHASSEE REGIONAL AP [ISIS], FL</td>
<td></td>
<td>2000-2019</td>
</tr>
<tr>
<td>726810</td>
<td>BOISE AIR TERMINAL [UO], ID</td>
<td></td>
<td>2000-2019</td>
</tr>
<tr>
<td>727640</td>
<td>BISMARCK MUNICIPAL ARPT [ISIS], ND</td>
<td></td>
<td>2000-2019</td>
</tr>
<tr>
<td>Station ID</td>
<td>Location Details</td>
<td>Solar Coordinates</td>
<td>Meteorological Coordinates</td>
</tr>
<tr>
<td>-----------</td>
<td>-----------------------------------</td>
<td>-------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>723650</td>
<td>ALBUQUERQUE INTL ARPT [ISIS], NM</td>
<td>Latitude: 35.04°</td>
<td>Latitude: 35.05°</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Longitude: -106.62°</td>
<td>Longitude: -106.8°</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Elevation: 1819 m</td>
<td>Elevation: 1818 m</td>
</tr>
<tr>
<td>723870</td>
<td>MERCURY DESERT ROCK AP [SURFRAD], NV</td>
<td>Latitude: 36.63°</td>
<td>Latitude: 36.617°</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Longitude: -116.02°</td>
<td>Longitude: -116.033°</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Elevation: 935 m</td>
<td>Elevation: 1006 m</td>
</tr>
<tr>
<td>725970</td>
<td>MEDFORD ROGUE VALLEY INTL AP [ASHLAND - UO], OR</td>
<td>Latitude: 42.19°</td>
<td>Latitude: 42.383°</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Longitude: -122.7°</td>
<td>Longitude: -122.667°</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Elevation: 585 m</td>
<td>Elevation: 396 m</td>
</tr>
<tr>
<td>726930</td>
<td>EUGENE MAHON SWEET ARPT [UO], OR</td>
<td>Latitude: 44.05°</td>
<td>Latitude: 44.117°</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Longitude: -123.07°</td>
<td>Longitude: -123.217°</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Elevation: 109 m</td>
<td>Elevation: 108 m</td>
</tr>
<tr>
<td>Location</td>
<td>Solar Coordinates</td>
<td>Meteorological Coordinates</td>
<td>Time Zone</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>---------------------------------------</td>
<td>----------------------------------------</td>
<td>-----------</td>
</tr>
<tr>
<td>AMARILLO INTERNATIONAL AP [CANYON - UT], TX</td>
<td>Latitude: 34.99° Longitude: -101.9° Elevation: 1088 m</td>
<td>Latitude: 35.217° Longitude: -101.7° Elevation: 1088 m</td>
<td>-6</td>
</tr>
<tr>
<td>SALT LAKE CITY INT’L ARPT [ISIS], UT</td>
<td>Latitude: 40.77° Longitude: -111.97° Elevation: 1288 m</td>
<td>Latitude: 40.783° Longitude: -111.987° Elevation: 1287 m</td>
<td>-7</td>
</tr>
<tr>
<td>WASHINGTON DC DULLES INT’L AR [STERLING - ISIS], VA</td>
<td>Latitude: 38.98° Longitude: -77.47° Elevation: 82 m</td>
<td>Latitude: 38.933° Longitude: -77.45° Elevation: 88 m</td>
<td>-5</td>
</tr>
</tbody>
</table>
### 726410
**MADISON DANE CO REGIONAL ARPT [ISIS], WI**

**Class I (with some measured)**

**Solar Coordinates**
- Latitude: 43.13°
- Longitude: -89.33°
- Elevation: 262 m

**Meteorological Coordinates**
- Latitude: 43.133°
- Longitude: -89.35°
- Elevation: 262 m

**Time Zone:** -6

---

### 726744
**ROCK SPRINGS ARPT [GREEN RIVER - UO], WY**

**Class I (with some measured)**

**Solar Coordinates**
- Latitude: 41.46°
- Longitude: -109.44°
- Elevation: 1000 m

**Meteorological Coordinates**
- Latitude: 41.6°
- Longitude: -109.067°
- Elevation: 2055 m

**Time Zone:** -7
REFERENCES


Akinoglu, B. G. (2008). Recent advances in the relations between bright sunshine hours and solar irradiation. In Modeling Solar Radiation at the Earth’s Surface (pp. 115-143). Springer Berlin Heidelberg.


