Modeling drought derivatives in arid regions: a case study in Qatar

A Thesis Submitted to the Committee on Graduate Studies in Partial Fulfillment of the Requirements for the Degree of Master of Science in the Faculty of Arts and Science

TRENT UNIVERSITY
Peterborough, Ontario, Canada

© Jayeong Paek 2016

Applied Modelling & Quantitative Methods M.Sc. Graduate Program

September 2016
ABSTRACT

Modeling drought derivatives in arid regions: a case study in Qatar

Jayeong Paek

We propose a stochastic weather model based on temperature, precipitation, humidity and wind speed for Qatar, as a representative arid region, in order to obtain simulated values for a drought index.

As a drought index, the Reconnaissance Drought Index (RDI) is commonly accepted in agriculture and is used to measure drought severity. It can be used to price weather derivatives to help farmers reduce financial losses from drought. RDI, which is the ratio of precipitation to evapotranspiration, is calculated by considering crop growth stages. The use of different crop coefficient value depending on the growth stage to calculate evapotranspiration can provide improved values for RDI. Additionally, six calculation methods for evapotranspiration using weather data are investigated to obtain accurate values for RDI.

Key Words: Stochastic weather models, Reconnaissance Drought Index (RDI), Evapotranspiration, Stochastic differential equations (SDEs), Mean reversion processes, Markov chains, crop coefficients.
Acknowledgements

First of all, I would like to appreciate my two supervisors, Prof. M. Pollanen and K. Abdella for giving me the opportunity to work with them. Their comments and advice led me to have successful results in my research work. They helped me so I can finish on time and only focus on my work by giving me a research grant.

I thank my colleagues at Trent University who we warm and welcoming. To all the faculty members of graduate studies and the Math department, I thank their services to make me get used to school easily.

I would also like to express my special thanks to Prof. Choi, Go and Kim at Chonnam National University and Prof. Kang at U of T who encourage me to study abroad. They continue helping me not to give up and to improve my research by providing valuable suggestion.

Lastly, I would like to show my gratitude to my family members and friends who always believe that I can do everything well. Their emotional support is a great motivation for being successful.
Contents

Abstract ii

Acknowledgements iii

1 Introduction 1

1.1 The objective of the thesis 1

1.2 Stochastic weather model 3

1.3 Outline of thesis 4

2 Background 7

2.1 Weather derivatives 7

2.2 Drought index 10

2.3 Evapotranspiration 12

2.4 Data description 14

2.5 Data preprocessing 15

2.6 R programming 17
3 Stochastic temperature model

3.1 Daily temperature model .............................................. 19
3.2 Trends and seasonality of temperature ............................. 21
3.3 Continuous temperature model with SDE ......................... 25
3.4 Parameter estimation .................................................... 27
  3.4.1 Least square ......................................................... 27
  3.4.2 Maximum likelihood estimation (MLE) .......................... 29
  3.4.3 Parameter estimation results ..................................... 32
3.5 Simulation ............................................................... 33

4 Stochastic precipitation model

4.1 Daily precipitation model .............................................. 35
4.2 Precipitation occurrence model ..................................... 37
  4.2.1 Markov chain ....................................................... 37
  4.2.2 Results of Markov chain ......................................... 40
4.3 Precipitation amount model ........................................... 43
  4.3.1 Probability distribution ........................................... 46
  4.3.2 Validation method for distribution ............................... 51
  4.3.3 Results .............................................................. 53
4.4 Simulation ............................................................... 57
  4.4.1 Simulation process ................................................. 57
5 Stochastic humidity and wind speed model 60

5.1 Stochastic humidity model 60

5.1.1 Trends and seasonality of humidity 62

5.1.2 Humidity model with SDE 66

5.1.3 Simulation result of humidity 67

5.2 Stochastic wind speed model 68

5.2.1 Wind speed model with Markov chain 70

5.2.2 Simulation results of wind speed 73

6 Reconnaissance Drought Index (RDI) 75

6.1 Evapotranspiration 76

6.1.1 Calculation methods for evapotranspiration 76

6.1.2 Comparison of calculation methods 79

6.1.3 Results of evapotranspiration 81

6.1.4 Comparison of evapotranspiration methods for RDI 89

6.2 Crop evapotranspiration 90

6.2.1 Crop coefficient 90

6.2.2 Results of crop evapotranspiration 91

6.3 Calculation of RDI for crops 97
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Daily temperature in Doha from 1983 to 2013 (°C)</td>
<td>22</td>
</tr>
<tr>
<td>3.2</td>
<td>Daily temperature in Doha for 1996 and 1997 (°C)</td>
<td>23</td>
</tr>
<tr>
<td>3.3</td>
<td>Empirical distribution of temperature in Doha (°C)</td>
<td>23</td>
</tr>
<tr>
<td>3.4</td>
<td>Comparison of observed and estimated temperature for Doha (°C)</td>
<td>25</td>
</tr>
<tr>
<td>4.1</td>
<td>Conditional probabilities of precipitation occurrence in Doha</td>
<td>42</td>
</tr>
<tr>
<td>4.2</td>
<td>Histogram of precipitation amount in Doha (mm)</td>
<td>44</td>
</tr>
<tr>
<td>4.3</td>
<td>Comparison of precipitation histogram and probability density</td>
<td>55</td>
</tr>
<tr>
<td>4.4</td>
<td>Comparison of rainy days from observation and simulation data</td>
<td>58</td>
</tr>
<tr>
<td>5.1</td>
<td>Comparison of temperature (°C) and humidity (%) for 1997 in Doha</td>
<td>62</td>
</tr>
<tr>
<td>5.2</td>
<td>Histogram of humidity in Doha (%)</td>
<td>63</td>
</tr>
<tr>
<td>5.3</td>
<td>Daily humidity in Doha from 1983 to 2013 (%)</td>
<td>64</td>
</tr>
<tr>
<td>5.4</td>
<td>Comparison of observed and estimated humidity (%)</td>
<td>65</td>
</tr>
<tr>
<td>5.5</td>
<td>Distribution of observation and simulation wind speed (m/s)</td>
<td>74</td>
</tr>
</tbody>
</table>
6.1 Plot of $ET_0$ in Doha (mm/day) ................................. 83
6.2 Comparative boxplots for $ET_0$ calculation methods ......... 84
6.4 Scatter plots of $ET_0$ .................................................. 88
6.5 Comparison of RDI in Doha from different methods of $ET_0$ 89
6.6 Comparative boxplots of $ET_c$ in Doha ............................. 94
6.7 Comparative boxplots of $ET_c$ in Doha by growing stages ...... 97
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Summary statistics of temperature in Doha (°C)</td>
<td>21</td>
</tr>
<tr>
<td>3.2</td>
<td>Parameter values of mean reversion process for temperature</td>
<td>32</td>
</tr>
<tr>
<td>3.2</td>
<td>Parameter values of mean reversion process for temperature</td>
<td>33</td>
</tr>
<tr>
<td>3.3</td>
<td>Comparison of summary statistics of temperature (°C)</td>
<td>34</td>
</tr>
<tr>
<td>4.1</td>
<td>Transition probability matrices for each month</td>
<td>40</td>
</tr>
<tr>
<td>4.1</td>
<td>Transition probability matrices for each month</td>
<td>41</td>
</tr>
<tr>
<td>4.2</td>
<td>Steady state probability vectors for precipitation occurrence in Doha</td>
<td>42</td>
</tr>
<tr>
<td>4.2</td>
<td>Steady state probability vectors for precipitation occurrence in Doha</td>
<td>43</td>
</tr>
<tr>
<td>4.3</td>
<td>Summary statistics for precipitation in Doha (mm)</td>
<td>45</td>
</tr>
<tr>
<td>4.4</td>
<td>Estimation of parameters in probability distributions</td>
<td>53</td>
</tr>
<tr>
<td>4.5</td>
<td>Results of KS test, AIC and BIC</td>
<td>56</td>
</tr>
<tr>
<td>4.6</td>
<td>Comparison of observation and simulation precipitation data</td>
<td>59</td>
</tr>
<tr>
<td>5.1</td>
<td>Summary statistics of humidity in Doha (%)</td>
<td>63</td>
</tr>
</tbody>
</table>
5.1 Summary statistics of humidity in Doha (%) .......................... 64
5.2 Parameters of mean reversion process for humidity .................. 66
5.3 Comparison of summary statistics of humidity (%) ..................... 67
5.3 Comparison of summary statistics of humidity (%) ..................... 68
5.4 Frequency of wind speed states in Doha ................................ 70
5.4 Frequency of wind speed states in Doha ................................ 71
5.5 Summary statistics of observation and simulation wind speed (m/s). 73
6.1 Summary statistics for $ET_0$ in Doha ..................................... 85
6.2 Statistics for the comparison of $ET_0$ methods ......................... 86
6.3 Growing periods and crop coefficients in Qatar (FAO 56, 1988) .... 90
6.3 Growing periods and crop coefficients in Qatar (FAO 56, 1988) .... 91
6.4 Summary statistics of $ET_0$ .................................................. 92
6.5 Mean of $ET_c$ by growing stage ............................................ 95
6.6 Summary statistics of RDI in Doha ....................................... 98
Glossary

**Crop coefficient**  The properties used in calculating the evapotranspiration for a crop. It has different values by crop type and growth stage, 13

**Drought index**  A measurement to quantify dryness. It is calculated by considering many weather factors, such as temperature, precipitation, humidity, wind speed and solar radiation, 7
<table>
<thead>
<tr>
<th><strong>Evapotranspiration</strong></th>
<th>The sum of evaporation and transpiration, which indicate water loss to the atmosphere from the soil and the plant, respectively. Since they occur simultaneously, both processes are treated together.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Growing stage</strong></td>
<td>The period in a crop life, separated by ground cover, crop height and leaf area. Typically, it is divided into four stages, initial, development, mid and end season.</td>
</tr>
<tr>
<td><strong>Reconnaissance Drought Index</strong></td>
<td>The ratio of cumulative precipitation and potential evapotranspiration. It is one of the drought indices commonly used in agriculture. In practice, actual evapotranspiration is used instead of using potential evapotranspiration.</td>
</tr>
</tbody>
</table>
Weather derivatives

Financial instruments used to protect financial losses related to adverse and unexpected weather. The main difference with typical derivatives is that underlying variables are indices that do not have financial values.
Chapter 1

Introduction

1.1 The objective of the thesis

It is a very well known fact that weather is crucial in deciding success and failure of yields in agriculture. Unfortunately, there is no way to control weather, or even predict it explicitly. Recently, one way to financially protect against production losses is to use weather derivatives. It has been considered a promising tool for hedging against weather (Barnett et al., 2008). Weather derivatives were introduced in the mid-1990’s and the market has been expanded rapidly, due to increasing weather variability. Weather derivatives originated from the energy industries to hedge risk caused by adverse weather conditions. As it is hard to predict demand for energy because of unexpected weather, energy companies need financial instruments to hedge against financial losses. In fact, most participants in weather derivatives market are
energy companies and there are many participants from other industries, such as transportation companies, construction companies, airports and even governments. Weather derivatives have been used as useful financial instruments to avoid financial losses from weather.

From the birth of agriculture, as a weather sensitive industry, farmers have been struggling with weather and trying to reduce financial losses derived from weather variability. By using weather derivatives, farmers can secure their profits. However, due to the difficulty in pricing the traded products and modeling the underlying weather variables, as well as the lack knowledge of the weather market and its benefits, weather derivatives are not widely used in agriculture. Especially, farmers, who do not know much about advanced financial markets and how they can protect themselves through weather derivatives, might be afraid to purchase weather derivatives.

Among the weather effects on agriculture, drought highly and directly influences production. Drought affects crop growth and leads to reduced production. The lack of water is a major risk to farmers in arid regions. Due to desertification, as dry regions become larger, the risk caused by drought will increase. Therefore, it is important work to examine the risk of drought in arid regions.

In agriculture, the Reconnaissance Drought Index (RDI) is considered as a good measurement to evaluate drought risk. Precipitation and evapotranspiration are important variables to decide RDI. Since evapotranspiration and precipitation are considered as moisture factors, they are desirable to be used to measure drought risk.
Evapotranspiration can be calculated from weather data and there are many calculation methods. These calculation methods were developed under certain climate conditions. Using proper methods is important to get an accurate RDI.

In this study, we make a stochastic weather model in terms of temperature, precipitation, humidity and wind speed. Through the stochastic weather model, we can provide simulation data that contains characteristics of weather including trends and seasonality. Simulation data from the stochastic weather model can be used to provide possible RDI values. Daily weather data from Qatar, representative of many arid countries, is used to make the weather model. By comparing RDI between observed and simulated data, we can see how well the simulation data works to provide information about drought risk. Furthermore, calculation methods for evapotranspiration are compared to find proper methods for arid regions.

1.2 Stochastic weather model

A deterministic model based on physical laws enables us to calculate the weather values at any time but it is not enough to explain weather variability. Weather is a time-dependent phenomenon but has many unknown factors unable to be explained from a deterministic model. Strong seasonality and trends in weather can be well explained by a deterministic model, however, unexpected weather conditions or extreme events, are difficult to reproduce from a deterministic model. For this reason,
it is desirable to use a stochastic model for weather to account for possible variability in those effects. The stochastic model approach provides a reasonable model to reproduce fluctuations and variability to improve model performance. Also it is better at reproducing the statistical properties of weather and replicates a variety of weather scenarios. This is because stochastic models can represent the uncertainty in the atmospheric processes by including random numbers, or noise terms, which are explained by probability distributions. In many previous studies, the stochastic approach to weather modeling has been used and showed good results.

1.3 Outline of thesis

This thesis is organized as follow.

In Chapter 2, we offer background on weather derivatives and drought indices. And weather data used to make stochastic weather model is explained. The preprocessing methods to fill out missing data are also discussed.

In Chapter 3, first of all we describe stochastic temperature models found in the previous literature. Stochastic differential equation (SDE) using Ornstein-Uhlenbeck is proposed for a daily temperature model. The two methods to estimate parameters in a SDE are explained. Finally summary statistics from simulation data are pre-
In Chapter 4, we make stochastic precipitation model by using a two-process. We apply a Markov chain model as an occurrence model and probability distributions are considered to decide the precipitation amount. In order to decide the proper probability distribution, a goodness of fit test is conducted.

Additionally, stochastic humidity and wind speed models are proposed in Chapter 5. In the same way as to temperature, a SDE is applied to model humidity. A Markov chain is applied for a wind speed model.

In Chapter 6, in order to calculate drought indices, six calculation methods to estimate evapotranspiration are examined. These methods are evaluated by calculating values of $R^2$, RMSE, MAE, MAXE, VE, CORR, EF and D by regression analysis. After we decide a proper method for evapotranspiration, RDI is calculated from observed and simulation data considering the growing stage.

The conclusion and future work are presented in Chapter 7. We summarize stochastic weather models for temperature, precipitation, humidity and wind speed and RDI. And we explain how this result can be used to help farmers in arid regions through weather derivatives. For the future work, we discuss how to improve perfor-
mance of stochastic weather models and problems found in the process of this study, such as preprocessing, weather modeling and calculation of RDI.
Chapter 2

Background

In this chapter, we introduce the three main concepts in this study: weather derivatives, drought indices and evapotranspiration. We explain the origin, usage and benefits of weather derivatives. Next, we focus on drought indices to measure drought intensity. We describe evapotranspiration, a part of calculating the drought index. Lastly, we explain the weather data used to make the stochastic weather models and data preprocessing methods for weather data.

2.1 Weather derivatives

Weather derivatives are financial tools that can be used to reduce financial risk and protect profits of organizations or individuals who are likely to be affected by adverse and unexpected weather. Participants in weather derivatives markets are insurance
and reinsurers, hedge funds, energy companies, pension funds, state governments, retailers, utility companies and so on. Government organizations that use weather derivatives can reduce their budget for contingencies caused from unexpected weather changes. Also, as many companies use weather derivatives to eliminate and hedge risk related to weather, they could be safe from the financial impacts of weather variability.

There is a big difference between weather derivatives and general derivatives. Underlying variables for weather derivatives, such as temperature, precipitation, humidity, snowfall and so on, have no financial value and they are not available to be traded and stored. Therefore, to decide the price of weather derivatives, indices associated with weather conditions are required. Through index values, the price of weather derivatives can be decided.

The Chicago Mercantile Exchange (CME), which is one of the largest option and future exchanges markets, offers future and option contracts based on weather, agricultural commodities, energy, foreign exchange, interest rates and metals. The CME provides diverse weather future and option contracts for weather in 10 cities around the world, 8 in the U.S. and 2 in Europe. In this market, Heating Degree Days (HDD) and Cooling Degree Days (CDD) are the most commonly used indices for temperature. In winter, temperature below 18 °C is measured and an associated index is referred to as HDD. In summer, temperature in excess of 18 °C is measured and an associated index is referred to as CDD.
There are three modeling approaches to pricing weather derivatives, Historical Burn Analysis (HBA), Index Modeling and Daily Modeling. As a classical approach, HBA uses the assumptions that the historical price of weather derivatives always provides good information for potential price. Thus the price of derivatives is calculated only based on historical price. It is very easy to calculate the price of weather derivatives because there is no need to apply distributions or models, but it is known to have large pricing errors (Jewson et al., 2005).

Another approach to pricing weather derivatives is Index Modeling. In index modeling, indices are directly modeled, for example, temperature indices such as HDD, CDD and the cumulative average temperature (CAT), are modeled for weather derivatives related to temperature. Therefore, we use simulated index values to estimate the price of weather derivatives. When index modeling is used, it is inevitable to lose information of extreme weather conditions. In addition, we need to find different models for each index. (Alexandridis, 2013)

Recent studies use the Daily Modeling approach that can provides simulation weather data. In Daily Modeling, we make a model for weather factors and then simulation data based on this model is used to calculate the price of weather derivatives. Therefore, it is important to have appropriate weather models in Daily Modeling. Once we have a weather model, any index values related to weather factors are available, so it can be used for weather derivatives based on different indices. In comparison to HBA and Index Modeling, it is much easier to connect with forecasting weather
data. Also as this method enables us to use complete weather data, it often provides more accurate weather derivatives prices than HBA (Jewson et al., 2005).

2.2 Drought index

Generally drought is considered a recurring natural phenomenon due to lack of water supply. Drought could be classified in many ways, depending on its type and characteristic, such as duration, magnitude, intensity, severity, geographic extent and frequency. From the World Meteorological Organization (1992) and American Meteorological Society (1997), drought has been defined as a “prolonged absence or marked deficiency of precipitation,” a “deficiency of precipitation that results in water shortage for some activity or for some group,” or a “period of abnormally dry weather sufficiently prolonged for the lack of precipitation to cause a serious hydrological imbalance”.

As we have many definitions, there are a variety of drought indices, more than 150, that are used for forecasting, monitoring and planning. Drought indices are developed and proposed for different application purposes. A drought index is better than using raw weather data to characterize drought levels and a useful tool to categorize intensity of drought intuitionally because it assimilates variables such as precipitation and evapotranspiration into a single numeric value. It makes it easy to establish strategies to reduce damage from drought by providing a good understanding of
drought. As long as it has a small value of drought index, it implies that it is dry. Through the drought index, drought could be classified as hyper-arid, arid, semi-arid, sub-humid and humid.

When assessing the severity of drought, evapotranspiration provides the most realistic information of water scarcity. Among the drought indices, Reconnaissance Drought Index (RDI) will be used in this study. Tsakiris et al (2007) states that RDI has many advantages for assessing drought. It is sensitive in changing climatic environments and flexible for different periods. Also it is preferred for use in agriculture to evaluate dryness in terms of growing crops because it has better an association with hydrological and agricultural droughts. RDI is the ratio of aggregated precipitation to potential evapotranspiration (PET), so a small value of RDI represents severe drought. It can be calculated as

$$RDI^i = \frac{\sum_{j=1}^{n} P_{ij}}{\sum_{j=1}^{n} PET_{ij}}$$

(2.1)

where $P_{ij}$ and $PET_{ij}$ are the precipitation and potential evapotranspiration (PET) for the $j$th month of the $i$th year and $n$ is a period in which we are interested. Following this equation, we can calculate RDI for any period in each year. Potential evapotranspiration (PET) is the evaporation that occurs where there is a sufficient water supply but in practice it is very hard to obtain. Instead of potential evapo-
transpiration (PET), actual evapotranspiration (ET) is used to calculate RDI. It is called adjusted RDI and written as

\[
RDI^i_{adj} = \frac{\sum_{j=1}^{n} P_{ij}}{\sum_{j=1}^{n} ET_{ij}},
\]

where \( ET_{ij} \) is the actual evapotranspiration (ET) for \( j \)th month of the \( i \)th year.

### 2.3 Evapotranspiration

The evapotranspiration is a combination of two processes that explain water loss to the atmosphere; evaporation and transpiration. Evaporation accounts for water movement from the soil to the atmosphere and transpiration explains water loss from plants. Evaporation and transpiration occur simultaneously, and are difficult to measure separately so we use evapotranspiration as a sum of two processes. Since the weather is the only factor to affect evapotranspiration, there is a calculation method using weather data. When estimating accurate evapotranspiration, temperature, humidity, wind speed and other weather parameters, such as solar radiation, pressure and so on, are required.

A number of calculation methods have been developed and proposed to estimate evapotranspiration from weather data. They have quite different forms and data
requirements because they were developed under certain weather conditions. For this reason, it is important to select proper calculation methods for evapotranspiration.

The FAO Penman-Monteith (PM) method is recommended as a standardized method from the Food and Agriculture Organization of the United Nations (FAO). However it requires us to use very specific weather variables, such as net radiation, soil heat flux density and psychometric constant, which are not easily obtained from general weather data. This is a reason why there are many methods for estimating evapotranspiration. From this calculation method, we can obtain evapotranspiration values, $ET_0$, which is called reference evapotranspiration. (Allen et al., 1998)

Once we obtain $ET_0$, crop evapotranspiration can be easily calculated. Most of the effects from various weather conditions are reflected in $ET_0$ so using crop coefficient we adjust $ET_0$ values reflecting the effect on crop type. By multiplying the crop coefficient and $ET_0$, we can calculate crop evapotranspiration, denoted by $ET_c$, that can be written as

$$ET_c = ET_0 \times k_c,$$

where $k_c$ is the crop coefficient. Crop coefficients are provided from FAO and these are used to calculate crop evapotranspiration. Crop coefficients are different not only by crop type but also growth stage. Therefore, different crops have different values of crop coefficients and while the crop is growing, it has also different crop coefficient values by growth stage.
2.4 Data description

In this study, daily weather data from Doha, Qatar is used. Qatar is located in the east of Saudi Arabia, on latitude 25.35 °N and longitude 51.03 °E. It has a dry climate with low precipitation and exceedingly hot and humid summers. During the winter months from December to February, the temperature remains above 10 °C and most of the precipitation occurs during this period. It receives a small amount of precipitation as shown from the Figure 4.2 and it is extremely unpredictable in time and space.

Weather data from 1983 to 2013 is used to make a stochastic weather model. To compare calculation methods for evapotranspiration, data over the period from 1985 to 2013 is used. It includes basic weather factors such as daily values of mean, minimum and maximum temperature (°C), precipitation (mm), relative humidity (%) and wind speed (m/s). More detailed weather factors for calculating evapotranspiration are included from 1985.

To ignore the effect of leap years, we subtract the last day, 29th, of February in the leap year. It gives us an easy model structure containing an equal number of days in each year. Thus, there are, 365, the same number of days in every year.
2.5 Data preprocessing

The most common problems found in weather data are missing and unreasonable or unreliable observations. Since weather data is collected and recorded from weather stations, the main reason of errors for weather data is problems within stations. Poor conditions of instruments or mistakes in the recording process can cause errors in weather data. Therefore, data should be corrected before it is used in many applications.

The unreasonable or unreliable observation in weather data, for example the minimum value is greater than maximum value or mean value is less than minimum value, can be found by using simple calculations. If we calculate difference between maximum and minimum values, the observation with opposite or rare sign of difference value should be checked.

The most common problem with the data is missing values and it is no exception in weather data. Dunis and Karalis (2003) suggest approaches to replace missing values for temperature. The simple way to replace missing values is to use temperature at the same day of the previous year. Another way is to use temperature at the same day from neighboring weather stations. Also he proposes an expectation maximization algorithm (EM algorithm), state space model, Kalman filter, neural network and principal component analysis (PCA), which are generally used in replacing missing values. Mraoua and Bari (2005) use PCA to replace missing temperature values and
conclude that it is a very efficient method for missing values under correlated data.

There is another elaborated method for replacing missing values in weather data. The Group Method of Data Handling (GMDH) is proposed to fill out missing values. GMDH is a method to use hierarchical polynomial regression networks developed to connect dependency between input and output. The basic idea of GMDH is to estimate output values by iterating processes of regression equations that contain subsets of input variables (Farrow, 1984). For polynomial equations, min and max temperature, solar radiation, precipitation and wind run were used for input variables and the commercial software package Model Quest was used (AbTech Corp., 1992-96) and each polynomial regression is able to select input variables automatically. Lebow et al. (1984) used the GMDH algorithm to predict seasonal precipitation and Lin et al. (1994) applied it to replace missing values for monthly mean temperature. Acock and Pachepsky (2000) explored the performance of GMDH to replace missing weather data from Mississippi for agriculture.

In the book written by Antonis et al. (2013), simple calculation methods for missing values in temperature are introduced. We need the average temperature on the same day over the \( N \) years of observations denoted by \( T_{Avg,t} \) and average temperature 7 days before and 7 days after the missing value denoted by \( T_{Avd,t} \). Finally, the average of \( T_{Avg,t} \) and \( T_{Avd,t} \) will be used to replace the missing value.

\[
T_{t,miss} = \frac{T_{Avg,t} + T_{Avd,t}}{2}, \tag{2.4}
\]
\[ T_{Avy,t} = \frac{1}{N} \sum_{yr=1}^{N} T_{t,yr}, \]  \hspace{1cm} (2.5)

\[ T_{Avd,t} = \frac{\left( \sum_{i=1}^{7} T_{t-i} + \sum_{i=1}^{7} T_{t+i} \right)}{14}. \]  \hspace{1cm} (2.6)

This method is very easy to calculate but it is difficult to apply when we have consecutive missing values.

In this study, missing values in the temperature data are replaced with the existing values at the same day in the previous year. This is because missing values found in our data are mostly consecutive and over a long period, for example the worst case is that there is no data for whole month. For precipitation, missing data values are considered as a non-rainy day because in Qatar there are not many rainy days.

### 2.6 R programming

In this study, R is used for the modelling and simulation work. R is a programming language that provides a variety of statistical analysis and graphical techniques. Robert Gentleman and Foss Ihaka from the University of Auckland developed R in 1993. The most powerful aspect of R is that, even though it is free, it is very easy to handle and transform data in R. Also it is very flexible to build own program code. Many packages for statistical analysis have been developed and offered so we can easily get and use. For this reason, R is the most used and popular program for statistical
analysis these days.
Chapter 3

Stochastic temperature model

In this chapter, a stochastic differential equation with an Ornstein-Uhlenbeck process is proposed in order to model daily temperature. The stochastic differential equation is solved using Itô’s Lemma. The parameters in the stochastic differential equation are estimated by the least square method and the maximum likelihood method based on a linear approach.

3.1 Daily temperature model

In early studies, daily temperature models were done using a discrete or continuous model. For the discrete model, a general autoregressive (AR) model was used for a daily temperature model in Carmona (1999), and it was compared to a mean-reverting discrete model in Moreno (2000). Moreno (2000) concluded that AR model is better
than the mean-reverting discrete model he proposed. Cao and Wei (2000) disputed the use of one-factor diffusion models, for which it is not possible to incorporate autocorrelation of temperature and suggested a k-lag autocorrelation model for daily temperature residuals. Cao and Wei (2000) considered unique characteristic of temperature to build a trend model for daily temperature, that is periodic cycles, moving around seasonal means, global warming and urban effects, autoregressive changes and higher volatility in winter than in summer.

The proposed continuous model includes a mean reverting term, which satisfies the characteristic of temperature that is moving around a seasonal mean. As a representative of mean reverting process, an Ornstein-Uhlenbeck (O-U) process called the Vasicek process is commonly used for stochastic temperature models. McIntyre and Doherty (1999) used an O-U process to model temperature at Heathrow airport in the UK and concluded that it shows a good fit to temperature data. Dornier and Queruel (2000) and Bhowan (2003) found that the expected value of the solution of a O-U process is not equal to the value where the O-U process should approach. To overcome this fact, he suggests an adjusted O-U process. Adjusted O-U process has been used to make a daily temperature model for Moroccan data (Mohammed and Driss, 2005).
3.2 Trends and seasonality of temperature

Before proceeding to modeling temperature, it is helpful to get an overall perspective for daily temperature. Seasonality is a very important component for a temperature model. In order to have better insight of temperature behavior, summary statistical values are calculated from Doha weather station. Mean, standard deviation, minimum and maximum of temperature by each month are presented in Table 3.1.

Table 3.1: Summary statistics of temperature in Doha (°C)

<table>
<thead>
<tr>
<th>Month</th>
<th>Mean</th>
<th>Min</th>
<th>Max</th>
<th>S.D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan</td>
<td>17.582</td>
<td>9.8</td>
<td>24.1</td>
<td>2.382</td>
</tr>
<tr>
<td>Feb</td>
<td>18.753</td>
<td>12.2</td>
<td>27.0</td>
<td>2.311</td>
</tr>
<tr>
<td>Mar</td>
<td>21.867</td>
<td>13.6</td>
<td>29.6</td>
<td>2.634</td>
</tr>
<tr>
<td>Apr</td>
<td>26.759</td>
<td>18.8</td>
<td>34.6</td>
<td>2.669</td>
</tr>
<tr>
<td>May</td>
<td>32.402</td>
<td>24.8</td>
<td>38.8</td>
<td>2.465</td>
</tr>
<tr>
<td>Jun</td>
<td>34.871</td>
<td>28.2</td>
<td>41.1</td>
<td>2.095</td>
</tr>
<tr>
<td>Jul</td>
<td>35.756</td>
<td>29.2</td>
<td>42.2</td>
<td>1.885</td>
</tr>
<tr>
<td>Aug</td>
<td>35.206</td>
<td>30.7</td>
<td>41.6</td>
<td>1.545</td>
</tr>
<tr>
<td>Sep</td>
<td>33.193</td>
<td>27.4</td>
<td>39.4</td>
<td>1.624</td>
</tr>
<tr>
<td>Oct</td>
<td>29.837</td>
<td>23.3</td>
<td>34.8</td>
<td>1.788</td>
</tr>
<tr>
<td>Nov</td>
<td>24.907</td>
<td>16.4</td>
<td>31.1</td>
<td>2.410</td>
</tr>
<tr>
<td>Dec</td>
<td>20.210</td>
<td>11.2</td>
<td>25.8</td>
<td>2.470</td>
</tr>
<tr>
<td>Total</td>
<td>27.670</td>
<td>9.800</td>
<td>42.20</td>
<td>6.897</td>
</tr>
</tbody>
</table>

The highest mean temperature is found in July, which is summer, and the lowest
mean temperature is obtained from January, which is winter. The standard deviation of temperature in winter is higher than summer. It is interesting that even in winter months, the temperature never goes below 0. So the mean temperature is always over $10^\circ C$.

![Image of temperature graph](image)

**Figure 3.1: Daily temperature in Doha from 1983 to 2013 ($^\circ C$)**

Daily temperature from 1983 to 2013 is plotted in Figure 3.1. It is apparent that temperature has strong seasonality. We can find a slight linear trend influenced by urbanization. Daily temperature from 1996 to 1997 is plotted in Figure 3.2. It is much easier to see seasonality and a periodic pattern.
The Figure 3.3 presents an empirical distribution of daily temperature. It indicates that temperature has a bimodal distribution in which there are two peaks, which are associated with summer and winter.

We can explain daily temperature by using a second order Fourier equation with
an added linear term. Seasonality can be described by sinusoids, as in a Fourier series and the linear trend is represented by \( a_0 + a_1 t \).

\[
DT(t) = a_0 + a_1 t + \sum_{i=1}^{2} b_i \cos(iwt) + \sum_{i=1}^{2} c_i \sin(iwt)
\]

\[
= a_0 + a_1 t + b_1 \cos(w t) + c_1 \sin(w t)
\]

where \( w = \frac{2\pi}{365} \), since temperature has seasonality with a period of 1 year. Using the method of least square, we find

\[
DT(t) = 26.3730 + 0.0002t - 3.4752 \cos(w t) - 8.6070 \cos(2w t)
\]

\[
-0.1227 \sin(w t) - 0.05792 \sin(2w t)
\]

The coefficient of the linear trend is 0.0002, which indicates that the daily temperature has risen slightly. The intercept, 26.3730, has a value close to the total average value of daily temperature.

Daily temperature estimated by the Fourier with linear trend model is presented with observed daily temperatures in Figure 3.4. For better visualization, only the first ten years of temperature are plotted in Figure 3.4. The line represents mean daily temperature obtained from the model and the dots are observed daily temperatures. It is interesting that much of the difference between estimated and observed temperature appears in winter and summer. It can be explained that temperature is much more erratic when temperature is very high and low.
3.3 Continuous temperature model with SDE

We propose to use a mean reverting O-U process as a daily temperature model. As suggested from Dornier and Queruel (2000) and Bhowan (2003), it is defined by

\[
dX_t = \left[ \lambda (\mu - X_t) + \frac{d\mu}{dt} \right] dt + \gamma dW_t, \tag{3.3}
\]

where \(X_t\) is the daily temperature, \(\lambda\) the speed of mean reversion, \(\mu\) the mean where the process reverts to, \(\gamma\) the volatility of the model and \(W_t\) is Wiener process which is normally distributed with mean of 0 and variance of \(t\). As long as it has large value of mean reversion, \(\lambda\), it means that it is very quickly pulled to back to its mean value \(\mu\).

We can solve this adjusted O-U process by using Itô’s Lemma. Itô’s Lemma
gives the correct expression for calculating differentials of composite functions which depend on Wiener process and also provides the SDE of a smooth function of a process satisfying an SDE. Let’s assume \( F = (\mu - X_t)e^{\lambda t} \) then we have

\[
\frac{\partial F}{\partial t} = \lambda e^{\lambda t} (\mu - X_t) + e^{\lambda t} \frac{d\mu}{dt},
\]

\[
\frac{\partial F}{\partial X_t} = -e^{\lambda t},
\]

\[
\frac{\partial^2 F}{\partial X_t^2} = 0.
\]

By Itô’s Lemma described in Appendix B,

\[
dF = \left[ \lambda e^{\lambda t} (\mu - X_t) + e^{\lambda t} \frac{d\mu}{dt} \right] dt - e^{\lambda t} dX_t
\]

\[
= \lambda e^{\lambda t} (\mu - X_t) dt + e^{\lambda t} d\mu - e^{\lambda t} \left[ \left\{ \lambda (\mu - X_t) + \frac{d\mu}{dt} \right\} dt + \gamma dW_t \right]
\]

\[
= -e^{\lambda t} \gamma dW_t.
\]

Thus, we have

\[
F(t) = F(0) + \int_0^t -e^{\lambda s} \gamma dW_s,
\]

and the solution of the adjusted O-U process is

\[
X_t = \mu - e^{\lambda t} (\mu + X_0) + e^{-\lambda t} \int_0^t e^{\lambda s} \gamma dW_s.
\]
From the above solution, the expected value of $X_t$ is

$$E(X_t|X_0) = X_0 e^{-\lambda t} + \mu (1 - e^{-\lambda t}) = \mu + (X_0 - \mu)e^{-\lambda t}. \tag{3.10}$$

Thus, we can prove that the adjusted O-U process approximately equals to the mean that the process should approach. In this process, $\lambda$, $\mu$ and $\gamma$ should be estimated from the data in hand. In the next section, we introduce two estimation methods for parameters in the SDE.

### 3.4 Parameter estimation

Least square regression and Maximum likelihood estimates are used to estimate parameters in the SDE.

#### 3.4.1 Least square

For the use of least square regression, it is assumed that consecutive observations have a linear relation with normally distributed error. The following linear equation is applied.

$$X_t = a + bX_{t-1} + \epsilon. \tag{3.11}$$
According to the solution of the SDE, we can find a relation between parameters in the SDE and linear equation. $\delta$ represents time step in the SDE.

$$a = \mu \left(1 - e^{-\lambda \delta}\right)$$

(3.12)

$$b = e^{-\lambda \delta}$$

(3.13)

$$sd(\epsilon) = \gamma \sqrt{\frac{1 - e^{-2\lambda \delta}}{2\lambda}}$$

(3.14)

Rewriting with respect to parameters in the SDE, we have

$$\lambda = -\frac{\ln b}{\delta},$$

(3.15)

$$\mu = \frac{a}{1 - b},$$

(3.16)

$$\gamma = sd(\epsilon) \sqrt{\frac{-2\ln b}{\delta (1 - b^2)}}.$$  

(3.17)

We can derive the parameters in the linear equation from the formulas of the sum of squares method:

$$a = \frac{S_y - bS_x}{n},$$

(3.18)

$$b = \frac{nS_{xy} - S_xS_y}{nS_{xx} - S_x^2},$$

(3.19)

$$sd(\epsilon) = \sqrt{\frac{nS_{yy} - S_y^2 - b(nS_{xy} - S_xS_y)}{n(n - 2)}},$$

(3.20)
where \( S_x = \sum_{i=1}^{n} X_{i-1}, \ S_y = \sum_{i=1}^{n} X_i, \ S_{xx} = \sum_{i=1}^{n} X_{i-1}^2, \ S_{xy} = \sum_{i=1}^{n} X_{i-1}X_i \) and \( S_{yy} = \sum_{i=1}^{n} X_i^2 \). Otherwise, we can easily find the parameters in the linear regression analysis.

### 3.4.2 Maximum likelihood estimation (MLE)

In statistics, MLE is a common method to estimate parameters using the likelihood function. MLE provides standard error, statistical test and other results useful for statistical inference. The likelihood function is a joint probability density function with respect to parameters, which are given by

\[
L(\theta; x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta), \quad i = 1, 2, \ldots, n.
\]  

(3.21)

Thus \( L(\theta|x) \) represents the likelihood of the parameter given the observed data and it is a function of \( \theta \). The Maximum likelihood estimator of \( \theta \) is the value that maximizes the likelihood function. It is the value that makes the observed data the most probable. Rather than maximizing the likelihood function, we use a log likelihood function. The logarithm is an increasing function so not only will it provide the same result when we use the likelihood function but also it is much easier to differentiate. The Maximum likelihood estimator can be calculated by solving the following partial differential equation:

\[
\frac{\partial \ln L(\theta|x)}{\partial \theta_j} = 0
\]

(3.22)
Applying the maximum likelihood estimation to the solution of the SDE, the conditional probability function of $X_t$ given $X_{t-1}$ is given by

$$f(X_t|X_{t-1}; \mu, \lambda, \hat{\gamma}) = \frac{1}{\sqrt{2\pi\hat{\gamma}^2}} \exp\left[-\frac{(X_t - X_{t-1}e^{-\lambda\delta} - \mu (1 - e^{-\lambda\delta}))^2}{2\hat{\gamma}^2}\right], \quad (3.23)$$

$$\hat{\gamma}^2 = \gamma^2 \frac{1 - e^{-2\lambda\delta}}{2\lambda}. \quad (3.24)$$

The log likelihood function is derived from the conditional probability function,

$$L(\mu, \lambda, \hat{\gamma}) = \sum_{t=1}^{n} \ln f(X_t|X_{t-1}; \mu, \lambda, \hat{\gamma})$$

$$= -\frac{n}{2} \ln 2\pi - n \ln \hat{\gamma} - \frac{1}{2\hat{\gamma}^2} \sum_{t=1}^{n} \left[X_t - X_{t-1}e^{-\lambda\delta} - \mu (1 - e^{-\lambda\delta})\right]^2. \quad (3.25)$$

The maximum likelihood estimator has a value that is satisfied when all partial derivatives with respect to each parameter are zero. Partial derivatives with respect to each parameter and solution are presented below.

$$\frac{\partial L(\mu, \lambda, \hat{\gamma})}{\partial \mu} = \frac{1}{\hat{\gamma}^2} \sum_{t=1}^{n} \left[X_t - X_{t-1}e^{-\lambda\delta} - \mu (1 - e^{-\lambda\delta})\right] (1 - e^{-\lambda\delta}) = 0, \quad (3.26)$$

$$\mu = \frac{\sum_{t=1}^{n} \left[X_t - X_{t-1}e^{-\lambda\delta}\right]}{n(1 - e^{-\lambda\delta})}, \quad (3.27)$$

$$\frac{\partial L(\mu, \lambda, \hat{\gamma})}{\partial \lambda} = -\frac{\delta e^{-\lambda\delta}}{\hat{\gamma}^2} \sum_{t=1}^{n} \left[(X_t - \mu)(X_{t-1} - \mu) - e^{-\lambda\delta}(X_{t-1} - \mu)^2\right] = 0, \quad (3.28)$$


\[
\lambda = -\frac{1}{\delta} \ln \frac{\sum_{t=1}^{n} (X_t - \mu)(X_{t-1} - \mu)}{\sum_{t=1}^{n} (X_{t-1} - \mu)^2},
\]

\[\frac{\partial L(\mu, \lambda, \hat{\gamma})}{\partial \hat{\gamma}} = -\frac{n}{\hat{\gamma}} + \frac{1}{\hat{\gamma}^3} \sum_{t=1}^{n} [X_t - X_{t-1}e^{-\lambda \delta} - \mu(1 - e^{-\lambda \delta})] = 0, \tag{3.30}\]

\[
\hat{\gamma}^2 = \frac{1}{n} \sum_{t=1}^{n} [X_t - X_{t-1}e^{-\lambda \delta} - \mu(1 - e^{-\lambda \delta})]^2. \tag{3.31}\]

However, this method has the problem that solutions of \(\mu\) and \(\lambda\) are dependent on each other. The solution of \(\mu\) is affected by \(\lambda\), likewise \(\lambda\) is affected by \(\mu\). Therefore, at least knowing one value of these two parameters is required to find \(\mu\) and \(\lambda\). In order to overcome this problem, a substitution of \(\lambda\) into \(\mu\) is applied. Once \(\mu\) and \(\lambda\) are decided, \(\gamma\) can be found easily by solution. Using the solution formulas of the sum of squares method, we can rewrite as follows.

\[
\mu = \frac{S_y - e^{-\lambda \delta} S_x}{n(1 - e^{-\lambda \delta})}, \tag{3.32}\]

\[
\mu = \frac{S_y S_{xx} - S_x S_{xy}}{n(S_{xx} - S_{xy})(S_x^2 - S_x S_y)}, \tag{3.33}\]

\[
\lambda = -\frac{1}{\delta} \ln \frac{S_{xy} - \mu S_x - \mu S_y + n\mu^2}{S_{xx} - 2\mu S_x + n\mu^2}, \tag{3.34}\]

\[
\hat{\gamma}^2 = \frac{1}{n} [S_{yy} - 2e^{-\lambda \delta} S_{xy} + e^{-2\lambda \delta} S_{xx}
- 2\mu(1 - e^{-\lambda \delta})(S_y - e^{-\lambda \delta} S_x) + n\mu^2(1 - e^{-\lambda \delta})^2], \tag{3.35}\]
\[ \gamma^2 = \gamma^2 \frac{2\lambda}{1 - e^{-2\lambda}}. \] (3.36)

### 3.4.3 Parameter estimation results

The parameters in the SDE are estimated for each month of the year. The mean values of the parameters in the SDE are calculated by month and presented in Table 3.2. The values of \( \mu \) and \( \lambda \) from the least square method and MLE are exactly the same but they provide different values of \( \gamma \). The differences of \( \gamma \) between least square method and MLE are very small. The highest value of \( \gamma \) is found in June and the smallest value is found in October. For a speed of mean reversion, August has a large value of \( \lambda \), which means it is drawn very strongly back to its mean value. Also we can find that the value of \( \mu \) is similar to the monthly mean of temperature.

<table>
<thead>
<tr>
<th>Month</th>
<th>( \mu )</th>
<th>( \lambda )</th>
<th>( \gamma ) (Least)</th>
<th>( \gamma ) (MLE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan</td>
<td>17.4780</td>
<td>0.4011</td>
<td>1.6294</td>
<td>1.5742</td>
</tr>
<tr>
<td>Feb</td>
<td>18.8917</td>
<td>0.4824</td>
<td>1.7278</td>
<td>1.6625</td>
</tr>
<tr>
<td>Mar</td>
<td>22.3762</td>
<td>0.4331</td>
<td>1.7986</td>
<td>1.7376</td>
</tr>
<tr>
<td>Apr</td>
<td>28.2374</td>
<td>0.2821</td>
<td>1.5822</td>
<td>1.5266</td>
</tr>
<tr>
<td>May</td>
<td>33.1215</td>
<td>0.3567</td>
<td>1.6389</td>
<td>1.5834</td>
</tr>
<tr>
<td>Jun</td>
<td>35.0124</td>
<td>0.6396</td>
<td>1.9190</td>
<td>1.8516</td>
</tr>
<tr>
<td>Jul</td>
<td>35.7646</td>
<td>0.6965</td>
<td>1.8501</td>
<td>1.7874</td>
</tr>
<tr>
<td>Aug</td>
<td>35.1229</td>
<td>0.9300</td>
<td>1.5149</td>
<td>1.4635</td>
</tr>
</tbody>
</table>
Table 3.2: Parameter values of mean reversion process for temperature

<table>
<thead>
<tr>
<th>Month</th>
<th>$\mu$</th>
<th>$\lambda$</th>
<th>$\gamma$ (Least)</th>
<th>$\gamma$ (MLE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sep</td>
<td>32.8898</td>
<td>0.5524</td>
<td>1.2789</td>
<td>1.2340</td>
</tr>
<tr>
<td>Oct</td>
<td>28.6960</td>
<td>0.2412</td>
<td>0.8825</td>
<td>0.0853</td>
</tr>
<tr>
<td>Nov</td>
<td>23.7035</td>
<td>0.1493</td>
<td>1.1662</td>
<td>1.1253</td>
</tr>
<tr>
<td>Dec</td>
<td>18.8716</td>
<td>0.3034</td>
<td>1.4249</td>
<td>1.3766</td>
</tr>
</tbody>
</table>

3.5 Simulation

We use the solution of the SDE to simulate daily temperatures. The following equation is used for generating simulation data for daily temperatures.

$$
X_t = \mu - (\mu - X_{t-1})e^{-\lambda \delta} + \gamma \sqrt{\frac{1 - e^{-2\lambda \delta}}{2\lambda}} N_{0,1}
$$

(3.37)

We simulate the data having the same length with observation data and calculate summary statistics to compare with observed temperature data. Summary statistics of observation and simulation data are summarized in Table 3.3. The mean of daily temperature from observation, 27.67, is similar to the one from simulation data, 27.57. Standard deviations from observation and simulation data, 6.8975 and 6.9036, are also close to each other. Additionally, when we find the 1st, 10th, 25th, 50th, 75th,
90th and 99th percentiles, we can see they are very similar. To sum up the results of summary statistics, we can conclude that temperature model with adjusted O-U process is good at reproducing the properties of temperature found in observation.

Table 3.3: Comparison of summary statistics of temperature (°C)

<table>
<thead>
<tr>
<th></th>
<th>Observation</th>
<th>Simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>27.67</td>
<td>27.57</td>
</tr>
<tr>
<td>SD</td>
<td>6.89756</td>
<td>6.903637</td>
</tr>
<tr>
<td>Min</td>
<td>9.80</td>
<td>11.48</td>
</tr>
<tr>
<td>Max</td>
<td>42.20</td>
<td>41.78</td>
</tr>
<tr>
<td>Percentile</td>
<td>1</td>
<td>14.22</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>17.90</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>21.40</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>28.70</td>
</tr>
<tr>
<td></td>
<td>75</td>
<td>33.90</td>
</tr>
<tr>
<td></td>
<td>90</td>
<td>35.90</td>
</tr>
<tr>
<td></td>
<td>99</td>
<td>38.60</td>
</tr>
</tbody>
</table>
Chapter 4

Stochastic precipitation model

In this chapter, we propose to use a Markov chain to model daily precipitation. A Markov chain is used to reproduce precipitation occurrence and a probability distribution is used to decide the precipitation amount on wet days. The two extreme probability distributions, GEVD and K4D, are proposed for the probability distribution of precipitation amount together with the Exponential, Lognormal and Gamma distributions. In order to determine the best probability distribution of the precipitation amount, Kolmogorov-Smirnov (KS test), AIC and BIC criteria are used.

4.1 Daily precipitation model

Musshoff et al. (2006) describes the characteristics of precipitation. First, the probability of precipitation occurrence follows a seasonal pattern. Second, the sequence
of wet and dry days is autoregressive. Third, the amount of precipitation on a rainy
day differs by season. Last, the volatility of the amount of precipitation occurrence
has a seasonal pattern.

Many stochastic models have been studied by considering the characteristics of
precipitation and two-process models are widely used. In the two-process model, a
Markov chain and probability distribution of precipitation amount are used. Williams
(1952) and Longley (1953) studied precipitation occurrence at Tel Aviv, Israel by
the Markov chain and showed that the number of rainy days is independent for
different months. In particular, a first-order Markov chain is proposed to be used to
model precipitation occurrence (Gabriel and Neumann, 1962; Katz, 1974; Chin, 1977;
Richardson, 1981; Wilks, 1999). However, the lack of memory length contained in a
first-order Markov chain was argued to be problematic so a higher-order Markov chain
was suggested to improve this deficiency (Dennett et al., 1983; Singh and Kripalani,
1986; Jones and Thornton, 1997; Wilks, 1999). In the recent research, a Markov chain
with jumps is proposed as an alternative model for precipitation by using weather data
from Chongqing, China. It shows better results for location with frequent extreme
precipitation events (Ahmet Göncü, 2011).
4.2 Precipitation occurrence model

As mentioned before, in two-processes model a Markov chain is used to generate precipitation occurrence. In this study, a first order Markov chain is applied to precipitation occurrence. Since the precipitation data can be separated into two types, zero and nonzero values.

4.2.1 Markov chain

Precipitation occurrence is modeled by a Markov chain. More precisely, a two-state, first-order Markov chain is widely used because it shows good performance to reproduce precipitation occurrence. The use of a higher-order Markov chain is able to improve model performance to reproduce precipitation occurrence. However it requires many calculations to estimate parameters whose number increases as the order increases. By considering numerical work, improvements of model performance through a higher-order is relatively low so a first-order is preferred to use.

The order number in a Markov chain represents the number of time steps in the past influencing the probability distribution of the present state. Thus, a first-order Markov chain has the assumptions that the probability of rain tomorrow depends only on whether today it is rainy or not. It can be explained by the Markov property,
as follows:

\[ P(X_{t+1} = s_{t+1} | X_t = s_t, X_{t-1} = s_{t-1}, \ldots, X_0 = s_0) = P(X_{t+1} = s_{t+1} | X_t = s_t), \quad (4.1) \]

where time \( t = \{0, 1, 2, \cdots, T\} \) and state space \( s = \{1, 2, 3, \cdots, S\} \). In a first-order Markov chain, the Markov property indicates that any future values are not relevant to past values. The only relevant value for the future values is the present value. In a Markov chain, a transition matrix is necessary to define a probability that each event occurs. It is composed of transition probabilities, which are conditional probabilities of future state \( j \) given state \( i \). The transition matrix, denoted \( P \), is given by

\[
P = \begin{bmatrix}
p_{11} & \cdots & p_{1j} \\
\vdots & \ddots & \vdots \\
p_{i1} & \cdots & p_{ij}
\end{bmatrix}
\quad \text{for } i, j \in S, \quad (4.2)
\]

where \( p_{ij} = P(X_{t+1} = j | X_t = i) \). The property of a transition matrix is that total sum of each row must equal 1.

\[
\sum_{j=1}^{S} P_{ij} = \sum_{j=1}^{S} P(X_{t+1} = j | X_t = i) = \sum_{j=1}^{S} P_{(X_t=i)}(X_{t+1} = j) = 1 \quad (4.3)
\]

Precipitation occurrence has two states: dry and wet. Therefore, the transition matrix is specified by two conditional probabilities, which are
\[ P_{dd} = P(\text{dry on day } t+1|\text{dry on day } t), \quad (4.4) \]

\[ P_{dw} = P(\text{wet on day } t+1|\text{dry on day } t), \quad (4.5) \]

\[ P_{wd} = P(\text{dry on day } t+1|\text{wet on day } t), \quad (4.6) \]

\[ P_{ww} = P(\text{wet on day } t+1|\text{wet on day } t). \quad (4.7) \]

Since there are only two states, transition probabilities at the same given state are complementary. So it is not necessary to estimate four transition probabilities, we can only estimate one of each pair of transition probabilities. For instance, the probability of a dry day following a dry day is calculated using the probability of wet day following the dry day, which is \( P_{dd} = 1 - P_{dw} \). The transitional matrix is defined as below.

\[
P = \begin{bmatrix} P_{dd} & P_{dw} \\
                 P_{wd} & P_{ww} \end{bmatrix}, \quad 0 \leq p_{ij} \leq 1 \quad i,j = \{d,w\}. \quad (4.8)
\]

Using a transition matrix, we can calculate the stationary state vector such that \( \pi = \pi P \). It implies a long-run relative frequency of precipitation occurrence and is satisfied with \( \sum_{i=1}^{S} \pi_i = 1 \), where \( \pi_i \geq 0 \) for all \( i \). Each element, \( \pi_i \), denotes the probability of being in state \( i \).

\[
\pi = \begin{pmatrix} \pi_d \\
                   \pi_w \end{pmatrix}, \quad (4.9)
\]
\[
\begin{pmatrix}
\pi_d & \pi_w
\end{pmatrix} = \begin{pmatrix}
\pi_d & \pi_w
\end{pmatrix} \begin{bmatrix}
1 - p_{dw} & p_{dw} \\
p_{wd} & 1 - p_{wd}
\end{bmatrix}.
\] (4.10)

In the case of a two-state first-order Markov chain, it can be calculated easily. Stationary probabilities \(\pi_d\) and \(\pi_w\) are given by

\[
\pi_d = \frac{p_{wd}}{p_{wd} + p_{dw}},
\] (4.11)

\[
\pi_w = \frac{p_{dw}}{p_{wd} + p_{dw}}.
\] (4.12)

### 4.2.2 Results of Markov chain

The transition probability matrixes for each month are given in Table 4.1. Since there are no rainy days in June and July for the entire Doha data set, the probability that dry day following dry day is 1. From November to April, the conditional probability of a wet day following a wet day is much greater than the months from May to October.

<table>
<thead>
<tr>
<th>Table 4.1: Transition probability matrices for each month</th>
</tr>
</thead>
</table>
| Jan \[
\begin{bmatrix}
0.977 & 0.023 \\
0.618 & 0.382
\end{bmatrix}
\] | Feb \[
\begin{bmatrix}
0.980 & 0.020 \\
0.680 & 0.320
\end{bmatrix}
\] |
| Feb \[
\begin{bmatrix}
0.974 & 0.026 \\
0.821 & 0.179
\end{bmatrix}
\] | Mar \[
\begin{bmatrix}
0.978 & 0.022 \\
0.905 & 0.095
\end{bmatrix}
\] |
Table 4.1: Transition probability matrices for each month

<table>
<thead>
<tr>
<th>Month</th>
<th>Transition Probability Matrix</th>
</tr>
</thead>
</table>
| May   | \[
|       | \begin{bmatrix} 0.993 & 0.007 \\ 1 & 0 \end{bmatrix} |
| Jun   | \[
|       | \begin{bmatrix} 1 & 0 \end{bmatrix} |
| Jul   | \[
|       | \begin{bmatrix} 1 & 0 \end{bmatrix} |
| Aug   | \[
|       | \begin{bmatrix} 0.998 & 0.001 \\ 1 & 0 \end{bmatrix} |
| Sep   | \[
|       | \begin{bmatrix} 0.999 & 0.001 \\ 1 & 0 \end{bmatrix} |
| Oct   | \[
|       | \begin{bmatrix} 0.998 & 0.002 \\ 1 & 0 \end{bmatrix} |
| Nov   | \[
|       | \begin{bmatrix} 0.985 & 0.015 \\ 0.619 & 0.381 \end{bmatrix} |
| Dec   | \[
|       | \begin{bmatrix} 0.973 & 0.027 \\ 0.650 & 0.350 \end{bmatrix} |

Figure 4.1 shows the monthly conditional probabilities of precipitation. According to Gabriel and Neumann (1962), due to the persistent nature of daily precipitation, occurrence conditional probability $p_{ww}$ should be larger than $p_{dw}$, so that $p_{dw} < \pi_w < p_{ww}$. In accordance with this result, the conditional probability of a wet day following a wet day, $p_{ww}$, is much greater than the conditional probability of a wet day following a dry day, $p_{dw}$, for every month. It means that a wet day is more likely to be followed
by a wet day. There is a great difference with regard to seasonal variation of monthly $p_{dw}$ and $p_{ww}$.

![Figure 4.1: Conditional probabilities of precipitation occurrence in Doha](image)

Based on the transition probability matrix, we can calculate the steady state probability vector and it is presented in Table 4.2. In both June and July, the steady state probability vector is equivalent to the transitional probability matrix because of no rainy day in these two months. As we can expect in an arid region, steady state probability for dry day is very high, which is close to 1.

<table>
<thead>
<tr>
<th>Month</th>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>Apr</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$[0.965 \ 0.035]$</td>
<td>$[0.971 \ 0.029]$</td>
<td>$[0.970 \ 0.030]$</td>
<td>$[0.976 \ 0.024]$</td>
</tr>
</tbody>
</table>

Table 4.2: Steady state probability vectors for precipitation occurrence in Doha
Table 4.2: Steady state probability vectors for precipitation occurrence in Doha

May  
\[
\begin{bmatrix}
0.993 & 0.007
\end{bmatrix}
\]

Jun  
\[
\begin{bmatrix}
1 & 0
\end{bmatrix}
\]

Jul  
\[
\begin{bmatrix}
1 & 0
\end{bmatrix}
\]

Aug  
\[
\begin{bmatrix}
0.999 & 0.001
\end{bmatrix}
\]

Sep  
\[
\begin{bmatrix}
0.999 & 0.001
\end{bmatrix}
\]

Oct  
\[
\begin{bmatrix}
0.998 & 0.002
\end{bmatrix}
\]

Nov  
\[
\begin{bmatrix}
0.976 & 0.024
\end{bmatrix}
\]

Dec  
\[
\begin{bmatrix}
0.960 & 0.040
\end{bmatrix}
\]

4.3 Precipitation amount model

Once we have the precipitation occurrence sequence, the next step is to determine the precipitation amount on wet days. On the contrary to the assumption of precipitation occurrence, the precipitation amount is considered independently. So probability distributions are used to decide precipitation amount. (Shamashad et al., 2005)

Many right skewed distributions with non-negative domains are investigated in the literature. This is because very small amounts of precipitation occur quite often. So the Exponential, Gamma and Lognormal distributions have been obtained as a
model of precipitation amount. Especially the Gamma distribution has been accepted as the most appropriate distribution in many previous works.

As we can see in the histogram of precipitation amount, Figure 4.2, most of the precipitation data is very small less than 10mm and there are few large values over 100mm. Based on this fact, the distribution of precipitation amount should be highly skewed to the right, which means it has a long right tail.

![Histogram of precipitation amount in Doha (mm)](image)

Figure 4.2: Histogram of precipitation amount in Doha (mm)

By accounting for the fact that there is an extreme amount of precipitation over 100mm, we expect that distribution with a thick tail can show good performance in deciding the precipitation amount. Two extreme distributions, general extreme value and 4-Kappa distributions are employed to model precipitation amount. Since the extreme distribution has a thick tail, it helps the model reproduces extremely large amount of precipitation in the simulation.
In this study, the probability distributions suggested in the previous research including the Exponential, Lognormal and Gamma distributions and two extreme distributions, the General Extreme Value and 4-Kappa distributions are used to model the precipitation amount. The parameters in the distributions are estimated by the maximum likelihood method (MLE), which is the most common method to find parameters in statistics.

Table 4.3: Summary statistics for precipitation in Doha (mm)

<table>
<thead>
<tr>
<th>Month</th>
<th>Mean</th>
<th>Max</th>
<th>SD</th>
<th>Total number of rainy days</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan</td>
<td>7.954</td>
<td>135.89</td>
<td>23.335</td>
<td>34</td>
</tr>
<tr>
<td>Feb</td>
<td>3.597</td>
<td>18.03</td>
<td>4.661</td>
<td>25</td>
</tr>
<tr>
<td>Mar</td>
<td>7.505</td>
<td>169.93</td>
<td>31.327</td>
<td>29</td>
</tr>
<tr>
<td>Apr</td>
<td>3.475</td>
<td>16.00</td>
<td>4.191</td>
<td>22</td>
</tr>
<tr>
<td>May</td>
<td>12.411</td>
<td>57.91</td>
<td>20.605</td>
<td>7</td>
</tr>
<tr>
<td>Jun</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>0</td>
</tr>
<tr>
<td>Jul</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>0</td>
</tr>
<tr>
<td>Aug</td>
<td>59.940</td>
<td>59.94</td>
<td>NA</td>
<td>1</td>
</tr>
<tr>
<td>Sep</td>
<td>71.880</td>
<td>71.88</td>
<td>NA</td>
<td>1</td>
</tr>
<tr>
<td>Oct</td>
<td>14.095</td>
<td>27.94</td>
<td>19.579</td>
<td>2</td>
</tr>
<tr>
<td>Nov</td>
<td>12.433</td>
<td>182.88</td>
<td>38.402</td>
<td>22</td>
</tr>
<tr>
<td>Dec</td>
<td>12.901</td>
<td>155.96</td>
<td>27.448</td>
<td>39</td>
</tr>
<tr>
<td>Total year</td>
<td>9.220</td>
<td>182.88</td>
<td>25.522</td>
<td>182</td>
</tr>
</tbody>
</table>
The summary statistics of precipitation for each month are shown in Table 4.3. The annual mean precipitation amount is 9.220 and standard deviation is 25.522. It means that Qatar has a very small amount of precipitation and it fluctuates a lot. The maximum precipitation for each month varies from 27mm to 182.88mm. There is only one rainy day in August and September so mean and maximum values are equal and we cannot calculate the standard deviation. Also from June to July, there is no recorded rainy day in the data. For 30 years, the total number of rain days is only 182 and 94.5% of wet days are found from November to April.

It makes sense that we have to find a probability distribution for each month but it does not work for our data because some months have very small data available. It can cause problems with parameter estimation. When we use MLE to estimate parameters, small size data can provide poor estimators. Thus, it would be better to use a whole year of data for the distribution of precipitation amount.

### 4.3.1 Probability distribution

In this section, the Exponential distribution, Lognormal distribution, Gamma distribution, General Extreme Value distribution (GEVD) and 4-Kappa distribution (K4D) are explained. They are applied to describe the distribution of precipitation amount.
(1) The Exponential distribution

The exponential distribution has been used to explain the distribution of waiting times between the events following poisson processes, where events occur independently and continuously with a constant rate. Because the Exponential distribution has a property of being memoryless, it is very useful to describe the time we have to wait before an event occurs regardless of how long we have already spent.

The exponential distribution has one parameter, $\lambda$, which is called the rate parameter, and whose probability density function is

$$ f(x; \lambda) = \lambda \exp(-\lambda x), \quad x \geq 0 $$

(4.13)

The distribution has mean $\mu = 1/\lambda$ and variance $\sigma^2 = 1/\lambda^2$. The maximum likelihood estimator of $\lambda$ is $1/\bar{x}$. Furthermore, the Exponential distribution is a special form of the Gamma distribution. The details are described in the Gamma distribution section.

(2) The Lognormal distribution

The Lognormal distribution is very closely connected to the normal distribution. It can be formed by a transformation of variable method from the normal distribution. In short, if $Y$, a random variable is normally distributed, then the random variable $X = \exp Y$ follows a Lognormal distribution. In other words, if random variable $Y$
has a Lognormal distribution, then the random variable $X = \ln Y$ has a normal distribution. For the relationship to the normal distribution, the Lognormal distribution is able to have only positive random values, which makes it possible to be used for a distribution of nonzero values.

The probability density function of the Lognormal distribution is given by:

$$f(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right),$$  \hspace{1cm} (4.14)

where the two parameters $\mu$ and $\sigma$ are, respectively, referred to as the location and scale parameters. The mean of the lognormal distribution is given by $\exp(\mu + \frac{1}{2}\sigma^2)$ and the variance by $(e^{\sigma^2} - 1) \exp(2\mu + \sigma^2)$. The maximum likelihood estimators of the two parameters respectively are:

$$\hat{\mu} = \frac{\sum_k \ln x_k}{n},$$  \hspace{1cm} (4.15)

$$\hat{\sigma}^2 = \frac{\sum_k (\ln x_k - \hat{\mu})^2}{n}.$$  \hspace{1cm} (4.16)

(3) The Gamma distribution

In the literature, the Gamma distribution is found to be a good probability distri-
bution for precipitation amount. The Gamma distribution has two parameters: the shape parameter $\alpha$ and the scale parameter $\beta$. The probability density function of the Gamma distribution is given by

$$f(x; \alpha, \beta) = \frac{\beta^{-\alpha} x^{\alpha-1} e^{-x/\beta}}{\Gamma(\alpha)}.$$  

$(4.17)$

$\Gamma(\alpha)$ is a gamma function evaluated at $\alpha$ and given by

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt.$$  

$(4.18)$

The mean of the Gamma distribution is $\alpha \beta$ and variance is $\alpha \beta^2$. As mentioned in the Exponential distribution section, the Gamma distribution can be described by a sum of exponential distributed values. The Gamma distribution can be described by a sum of mutually independent random variables having the Exponential distribution. The Gamma distribution with $\alpha = 1$ and $\beta$ is as same as in the Exponential distribution having $\lambda = \beta$.

(4) The General Extreme Value distribution (GEVD)

The Gumbel, Fréchet and Weibull distributions are typically used to describe extreme value behavior and it is usual to adopt one of the three in applications. The single form has been developed for better analysis by combining these three distributions
and it is called the General Extreme Value distribution (Stuart Coles, 2001). Its
probability density function is defined by

\[
G(x) = \exp \left\{ - \left[ 1 + \xi \left( \frac{x - \mu}{\sigma} \right) \right]^{-1/\xi} \right\}, \tag{4.19}
\]

where \( \{x : 1 + \xi(z - \mu)/\sigma \} > 0 \) and \(-\infty < \mu < \infty, \sigma > 0, -\infty < \xi < \infty. \) It has
three parameters; the location parameter \( \mu; \) the scale parameter \( \sigma; \) and the shape
parameter \( \xi. \) Depending on the different parameter values, it can take the form of a
Gumbel, Fréchet or Weibull distributions.

(5) 4-Kapppa distribution (K4D)

The 4-Kappa distribution(K4D) is widely used in extreme data analysis together with
GEVD. Hosking(1994) who pioneered the Kappa distribution has investigated appli-
cations of kappa in many fields and methods for parameter estimate. The probability
density function of K4D is given by

\[
f(x; \xi, \alpha, h, k) = \alpha^{-1}[1 - ky]^{(1/k)-1}[F(x)]^{1-h}, \quad k \neq 0 \ h \neq 0. \tag{4.20}
\]

Furthermore, it can obtain many different distributions depending on parameter
value. When it comes to the parameter \( h, \) if it has value 1, K4D will become the Gen-
eral Pareto distribution; in case of 0 it will take GEVD and with \( h = -1 \) it becomes
the General Logistic distribution. When $h$ and $k$ are 0 then Gumble distribution can be obtained.

### 4.3.2 Validation method for distribution

It is necessary to choose a proper distribution for precipitation amount in precipitation amount modeling. In statistics, it is called Goodness-of-fit testing and there are many methods to validate model to see how well it fits data. Three methods are used in this study: The Kolmogorov-Smirnov test, Akaike information criterion and Bayesian information criterion.

#### (1) Kolmogorov-Smirnov test (KS test)

The Kolmogorov-Smirnov test (KS test) is a nonparametric test for deciding if observed data comes from a specific distribution. The KS test is conducted based on the empirical distribution function. This test measures differences between the empirical distribution of observed data and the cumulative distribution of the reference distribution. Among these differences, the greatest value is going to be KS statistic, which is denoted D. The advantage of the KS test is that there is no required assumption for the distribution of the observed data. As with the results of hypothesis test, we are provided statistic D and p-values to make a decision.
(2) Akaike information criterion (AIC-Hirotugu Akaike, 1974)

The AIC was developed by Hirotugu Akaike and first suggested to evaluate performance of linear regression models. The basic idea of AIC is to find a model, which minimizes the loss of information from given data. He created the AIC value based on a likelihood function, which is defined by:

\[
AIC = -2 \ln(L) + 2K
\]  

(4.21)

where \( L \) is the log-likelihood function and \( K \) is the number of parameters in the model. It contains a penalty term \( K \) with respect to the numbers of parameters. Although using many parameters in a model improves model performance, it can cause over-fitting problems. The model with the smallest AIC value is selected as a best model for the given data.

(3) Bayesian information criterion (BIC-Schwarz, 1978)

The BIC is closely related to AIC in terms of using the likelihood function. However, BIC has a different penalty term compared to AIC. It is also based on the likelihood function and has a penalty term according to not only the number of parameters but also data size. The BIC is given by Gideon E. Schwarz in 1979 and defined by

\[
BIC = -2 \ln(L) + K \ln(n)
\]  

(4.22)
where $L$ is log-likelihood function, $K$ is the number of parameters and $n$ is the size of data. As with AIC, the model with the smallest BIC values is preferred to be a good model.

### 4.3.3 Results

MLE is used to estimate parameters in five probability distributions and results are shown in Table 4.4.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Estimation of parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential</td>
<td>$\hat{\mu} = 0.108$</td>
</tr>
<tr>
<td>Lognormal</td>
<td>$\hat{\mu} = 0.834 \quad \hat{\sigma} = 1.510$</td>
</tr>
<tr>
<td>Gamma</td>
<td>$\hat{\alpha} = 0.463 \quad \hat{\beta} = 0.050$</td>
</tr>
<tr>
<td>GEVD</td>
<td>$\hat{\mu} = 1.060 \quad \hat{\alpha} = 1.322 \quad \hat{\xi} = 1.364$</td>
</tr>
<tr>
<td>K4D</td>
<td>$\hat{\xi} = 1.225 \quad \hat{\xi} = 1.424 \quad \hat{h} = -0.375 \quad \hat{k} = -1.329$</td>
</tr>
</tbody>
</table>

The Figure 4.3 illustrates the histogram of precipitation amounts with fitted probability distribution respectively. The dashed line represents the probability density function for each distribution with parameters estimated by MLE. We can find that all distributions are strongly right skewed.
CHAPTER 4. STOCHASTIC PRECIPITATION MODEL

(a) Exponential

(b) Lognormal

(c) Gamma
Figure 4.3: Comparison of precipitation histogram and probability density

It is hard to figure out which distribution shows the best fit to precipitation amount by only comparing figures. As mentioned before, using the three model validation methods we can compare the performance of the distributions statistically. The KS test provides p-values as a results and AIC and BIC have statistic values from the equation made up of the likelihood function and penalty terms associated with the number of parameters in the distribution and the number of data used. In the KS
test, the significance level of 0.05 is used to determine adaptation of distribution. For instance, if the p-value provided from KS test is less than the significance level, we conclude that the given distribution is not proper for the data in hand; otherwise we accept the use of given distribution. For both AIC and BIC, distribution with the smallest value is recommended to be used. The results of KS test, AIC and BIC are presented in Table 4.5.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>KS test (p-value)</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential</td>
<td>2.2e-16</td>
<td>1174.589</td>
<td>1177.793</td>
</tr>
<tr>
<td>Lognormal</td>
<td>9.937e-04</td>
<td>973.921</td>
<td>980.329</td>
</tr>
<tr>
<td>Gamma</td>
<td>1.467e-06</td>
<td>1075.677</td>
<td>1082.085</td>
</tr>
<tr>
<td>GEVD</td>
<td>1.467e-06</td>
<td>953.895</td>
<td>963.507</td>
</tr>
<tr>
<td>K4D</td>
<td>0.0929</td>
<td>999.214</td>
<td>1008.03</td>
</tr>
</tbody>
</table>

The p-value of all distributions presented in Table 4.5 are significantly very small, less than the significance level of 0.05, except the one from K4D. It indicates that only K4D is a significant distribution resulted from the KS test. With regard to AIC and BIC, GEVD has the smallest value compared to the other distributions. To sum up these two consequences, two extreme distributions, K4D and GEVD, resulted in proper distributions for precipitation amount.
4.4 Simulation

4.4.1 Simulation process

The following steps are required to generate the precipitation occurrence sequence and determine precipitation amount on wet days. The first three steps are for generating precipitation occurrence using a Markov chain and last step is for generating precipitation amount using the probability distribution.

**Step 1.** Find the transitional matrix for each month and make sure that sum of each row should be equal to 1. Calculate the stationary state probability vector for a long period.

**Step 2.** Generate a random number from a uniform distribution with interval between 0 and 1 then compare the random number to the stationary probabilities. If the random number is less than $\pi_d$ then it is dry otherwise it is wet.

**Step 3.** Repeat the two previous steps until we get the same number of days for each month.

**Step 4.** Once we get the precipitation occurrence sequence, a random value from the probability distribution fitted to precipitation amount is assigned to the
precipitation amount on the wet days.

4.4.2 Simulation results of precipitation

A 30-year simulation data was generated using the stochastic precipitation model with a Markov chain. The number of wet days in the observation and simulation data is presented in Figure 4.4. The number of wet days between the observation and simulation data is very similar. It indicates that a first order Markov chain works well to generate precipitation occurrence. From the simulation data, wet days in January, March and August are generated more than observation data. Whereas we can find that wet days in February, April, May, October, November and December are generated less than observation data.

![Figure 4.4: Comparison of rainy days from observation and simulation data](image)

In order to compare precipitation amounts between observation and simulation
data, percentiles and mean are calculated. Minimum, 25-percentiles (Q1), median (Q2), mean, 75-percentiles (Q3) and maximum of simulation precipitation data are presented in Table 4.6. It is obvious that a right skewed distribution is appropriate for distribution of precipitation amount since the median of the observation precipitation amount is smaller than the mean. In both simulations using GEVD and K4D, the median is smaller than the mean. Most of percentiles from the two simulations are very similar to those from observation data but the maximum value from K4D is much closer to the one from observation data. For this reason, K4D is better for the precipitation amount distribution rather than GEVD.

Table 4.6: Comparison of observation and simulation precipitation data

<table>
<thead>
<tr>
<th></th>
<th>Observation</th>
<th>Simulation 1 (GEVD)</th>
<th>Simulation 2 (K4D)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum</td>
<td>0.25</td>
<td>0.17</td>
<td>0.19</td>
</tr>
<tr>
<td>Q1</td>
<td>0.76</td>
<td>0.63</td>
<td>0.67</td>
</tr>
<tr>
<td>Q2 (Median)</td>
<td>1.78</td>
<td>1.50</td>
<td>1.53</td>
</tr>
<tr>
<td>Mean</td>
<td>9.22</td>
<td>6.92</td>
<td>8.68</td>
</tr>
<tr>
<td>Q3</td>
<td>6.10</td>
<td>5.24</td>
<td>4.71</td>
</tr>
<tr>
<td>Maximum</td>
<td>182.88</td>
<td>96.99</td>
<td>186.04</td>
</tr>
</tbody>
</table>
Chapter 5

Stochastic humidity and wind speed model

In this chapter, stochastic models for humidity and wind speed are proposed. Mean reverting O-U process, which is the same approach used with the temperature model is applied to the humidity model. Also a first-order Markov chain is applied to the wind speed model.

5.1 Stochastic humidity model

Relative humidity affects how much people feel hot and cold temperatures. For example, if the air is very humid in summer then we feel much hotter than the actual temperature. In other words, if the relative humidity is low in summer, we can feel
cooler than the actual temperature. This is because of the process of evaporation, which is called sweating in humans. When the air is dry, which means low relative humidity, and the temperature is high, evaporating from our body works easily and quickly to make us cool off. For this reason, under the same temperature we feel different depending on relative humidity.

Similarly, it can be applied to growing crops. If relative humidity is high, the amount of water required to grow crops might be lessened because evaporation and transpiration from crops work more slowly.

The relative humidity (RH) is defined as the ratio of the actual water vapor pressure \( e \) to the saturation water vapor pressure \( e_s \).

\[
RH(\%) = 100 \frac{e}{e_s} \tag{5.1}
\]

Based on the fact that RH and dew point are the indicators of the amount of air moisture, approximation of RH using dew point can be defined as below.

\[
RH(\%) \approx 100 - 5(t - t_d) \tag{5.2}
\]

where \( t \) is temperature and \( t_d \) is dew point. The relationship between RH and dew point is non-linear but when RH is more than 50\%, it becomes nearly linear. Therefore, it is reasonable that a model of humidity should be explained by the relationship
with temperature or dew point. However, a humidity model based on a meteorological explanation makes the model complicated and requires the use of other weather factors, such as temperature or dew point. However, in this research, we first try to make a humidity model with respect to time, not considering any relationship with other weather factors.

5.1.1 Trends and seasonality of humidity

Daily temperature and RH in 1997 are presented together in Figure 5.1 to make it easy to compare them. As you can see from Figure 5.1, RH shows inverse behavior to temperature and the volatility of RH is much bigger than for temperature.

![Figure 5.1: Comparison of temperature (°C) and humidity (%) for 1997 in Doha](image)

The Figure 5.2 presents empirical distribution of RH. Different than for temperature, it shows a unimodal distribution.
Summary statistics of RH are presented in Table 5.1. We can see RH has a large range, varying from 8 to 95. The interesting fact found from the maximum humidity is that regardless of seasonality, maximum humidity is always over 80. It is quite different than temperature, where the maximum temperature also shows a seasonal pattern.

Table 5.1: Summary statistics of humidity in Doha (%)
Table 5.1: Summary statistics of humidity in Doha (%)

<table>
<thead>
<tr>
<th>Month</th>
<th>Mean</th>
<th>Min</th>
<th>Max</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jul</td>
<td>43.37461</td>
<td>13</td>
<td>82</td>
<td>16.318035</td>
</tr>
<tr>
<td>Aug</td>
<td>54.58793</td>
<td>14</td>
<td>84</td>
<td>13.899014</td>
</tr>
<tr>
<td>Sep</td>
<td>32.043</td>
<td>17</td>
<td>80</td>
<td>11.484558</td>
</tr>
<tr>
<td>Oct</td>
<td>61.811</td>
<td>29</td>
<td>85</td>
<td>9.235284</td>
</tr>
<tr>
<td>Nov</td>
<td>62.04624</td>
<td>32</td>
<td>86</td>
<td>8.651923</td>
</tr>
<tr>
<td>Dec</td>
<td>67.65765</td>
<td>33</td>
<td>94</td>
<td>9.588968</td>
</tr>
<tr>
<td>Total</td>
<td>54.04127</td>
<td>8</td>
<td>95</td>
<td>16.26033</td>
</tr>
</tbody>
</table>

Similar to temperature, RH shows seasonality but it is hard to find a linear trend in RH. Simply, it can be explained by applying Fourier series.

Figure 5.3: Daily humidity in Doha from 1983 to 2013 (%)
A second order Fourier equation is applied to explain seasonality of RH. The equation with parameters estimated by using least square is given by

\[
DRH(t) = 54.04102 + 0.02909 \cos(wt) + 13.67946 \cos(2wt)
\]

\[
-0.01251 \sin(wt) - 4.37344 \sin(2wt)
\]

(5.3)

where \( w = 2\pi/365 \). RH estimated by a second order Fourier series is presented with observed RH in Figure 5.4 over the period of ten years. The line represents RH obtained from a Fourier series and dots represent observed RH. We can see that RH fluctuates from the estimated value a lot.

Figure 5.4: Comparison of observed and estimated humidity (%)
5.1.2 Humidity model with SDE

A mean reversion O-U process is applied to model daily humidity data. The parameters estimated by using least square and MLE are listed in Table 5.2. As mentioned in temperature section, the parameters estimated by the two methods have equal values of speed of mean reversion $\lambda$ and mean $\mu$ but there is a small difference in a value of volatility $\gamma$. As we can expect, the value of volatility is larger values than those from temperature model.

Table 5.2: Parameters of mean reversion process for humidity

<table>
<thead>
<tr>
<th>Month</th>
<th>$\mu$</th>
<th>$\lambda$</th>
<th>$\gamma$(Least)</th>
<th>$\gamma$(MLE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan</td>
<td>67.4488</td>
<td>0.6768</td>
<td>11.4293</td>
<td>11.0417</td>
</tr>
<tr>
<td>Feb</td>
<td>64.0371</td>
<td>0.9049</td>
<td>14.8098</td>
<td>14.2507</td>
</tr>
<tr>
<td>Mar</td>
<td>54.9218</td>
<td>0.7767</td>
<td>15.1913</td>
<td>14.6762</td>
</tr>
<tr>
<td>Apr</td>
<td>46.4332</td>
<td>0.5920</td>
<td>12.7813</td>
<td>12.3327</td>
</tr>
<tr>
<td>May</td>
<td>36.3553</td>
<td>0.7387</td>
<td>12.2807</td>
<td>11.8643</td>
</tr>
<tr>
<td>Jun</td>
<td>35.3932</td>
<td>0.7081</td>
<td>14.4253</td>
<td>13.9190</td>
</tr>
<tr>
<td>Jul</td>
<td>45.2631</td>
<td>0.5689</td>
<td>15.4916</td>
<td>14.9663</td>
</tr>
<tr>
<td>Aug</td>
<td>55.2619</td>
<td>0.7355</td>
<td>14.2856</td>
<td>13.8012</td>
</tr>
<tr>
<td>Sep</td>
<td>56.2345</td>
<td>1.1614</td>
<td>15.9156</td>
<td>15.3570</td>
</tr>
<tr>
<td>Oct</td>
<td>58.7774</td>
<td>0.9829</td>
<td>11.9194</td>
<td>11.5152</td>
</tr>
<tr>
<td>Nov</td>
<td>62.1151</td>
<td>1.1193</td>
<td>11.0523</td>
<td>10.6644</td>
</tr>
<tr>
<td>Dec</td>
<td>67.9820</td>
<td>0.8075</td>
<td>10.6235</td>
<td>10.2633</td>
</tr>
</tbody>
</table>
5.1.3 Simulation result of humidity

The solution of the SDE is used to simulate humidity. However, we find a significant problem in simulating humidity. The problem is that the values provided from the model are possibly outside the range of possible values. Since humidity has non-negative values between 0 and 100, the values out of the range are not proper. In order to overcome this, we repeat the simulation process until we get a proper value within the range.

Summary statistics from observation and simulation data are summarized in Table 5.3. The mean of humidity in simulation is very close to observation. In addition, percentiles in simulation are very similar to each other. Humidity simulation shows smaller standard deviation than observation. The big difference is found in the minimum values. In observation, minimum humidity is recorded 8 but 2.252 is provided in simulation.

<table>
<thead>
<tr>
<th></th>
<th>Observation</th>
<th>Simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>54.01</td>
<td>54.210</td>
</tr>
<tr>
<td>SD</td>
<td>16.2691</td>
<td>14.92881</td>
</tr>
<tr>
<td>Min</td>
<td>8.00</td>
<td>2.252</td>
</tr>
<tr>
<td>Max</td>
<td>95.00</td>
<td>98.900</td>
</tr>
</tbody>
</table>
From the results, a humidity model with mean reverting O-U process shows good performance of reproducing statistical values. However, it is not enough to be a complete model because we never consider the relationship between humidity and other weather factors in the model.

### 5.2 Stochastic wind speed model

Wind is considered as one of the promising renewable energy sources of the future. As demand increases for renewable energy source to replace fossil fuels, wind has been noted from energy planners and producers in wind farms. Understanding the statistical parameters of wind speed are important to construct a proper wind energy
Many probability distributions are suggested to fit wind speed. These distributions include the Weibull, Lognormal, Inverse Gaussian, Gamma, Erlang and Gumbel distributions (Luna, 1974; Jason, 2001; Brano, 2011). The Weibull distribution is the most popular distribution to explain wind speed. Using probability distribution to model wind speed requires the assumption that consecutive wind speeds are independent. However it cannot be acceptable for a model with short interval wind speeds such as hourly or daily.

Another approaches to model wind speed are Autoregressive processes, Markov chains, Neural Networks methods and Wavelet-based approaches. Autoregressive processes and Markov chains are suggested to consider dependency in time series (Jones, 1986 and Kirchoff, 1989). Shamshad et al. (2005) compared first and second order Markov chains using hourly wind speed data from Malaysian Meteorological stations and conclude that second order Markov chain shows improved results. Aksoy et al. (2004) compared probability distribution methods including the Normal and Weibull distributions, first and second order autoregressive processes, first and second order Markov chains and wavelet-based approaches using hourly wind speed from a Southeastern Anatolian city over four years. Statistical values between observed and simulated data from the different models are provided as results. Wavelet-based approach results show the best performance for preserving wind speed behavior in time series.
5.2.1 Wind speed model with Markov chain

A first-order Markov chain is used for a wind speed model. As it is already described in the previous chapter on the precipitation model, using a first-order Markov chain means that we only consider wind speed today to decide the wind speed tomorrow. The difference with the precipitation and wind speed is that we use many more states than two in a wind speed model.

Minimum and maximum wind speed is used to decide the number of states in a Markov chain. Minimum and maximum values of observed wind speed are 2.4 m/s and 39.1 m/s, respectively. Thus, 38 states are chosen to divide wind speed and each state is designed to have 1 m/s width. The first and last states start with 2 m/s and 39 m/s, respectively. Table 5.4 presents frequency of wind speed in each state.

Table 5.4: Frequency of wind speed states in Doha

<table>
<thead>
<tr>
<th>State</th>
<th>Frequency</th>
<th>State</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>139</td>
<td>20</td>
<td>300</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>21</td>
<td>210</td>
</tr>
<tr>
<td>3</td>
<td>56</td>
<td>22</td>
<td>185</td>
</tr>
<tr>
<td>4</td>
<td>157</td>
<td>23</td>
<td>162</td>
</tr>
<tr>
<td>5</td>
<td>408</td>
<td>24</td>
<td>146</td>
</tr>
<tr>
<td>6</td>
<td>648</td>
<td>25</td>
<td>134</td>
</tr>
<tr>
<td>7</td>
<td>862</td>
<td>26</td>
<td>32</td>
</tr>
<tr>
<td>8</td>
<td>1083</td>
<td>27</td>
<td>72</td>
</tr>
</tbody>
</table>
Table 5.4: Frequency of wind speed states in Doha

<table>
<thead>
<tr>
<th>State</th>
<th>Frequency</th>
<th>State</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>955</td>
<td>28</td>
<td>64</td>
</tr>
<tr>
<td>10</td>
<td>938</td>
<td>29</td>
<td>35</td>
</tr>
<tr>
<td>11</td>
<td>708</td>
<td>30</td>
<td>50</td>
</tr>
<tr>
<td>12</td>
<td>669</td>
<td>31</td>
<td>23</td>
</tr>
<tr>
<td>13</td>
<td>672</td>
<td>32</td>
<td>11</td>
</tr>
<tr>
<td>14</td>
<td>519</td>
<td>33</td>
<td>14</td>
</tr>
<tr>
<td>15</td>
<td>537</td>
<td>34</td>
<td>9</td>
</tr>
<tr>
<td>16</td>
<td>385</td>
<td>35</td>
<td>6</td>
</tr>
<tr>
<td>17</td>
<td>368</td>
<td>36</td>
<td>5</td>
</tr>
<tr>
<td>18</td>
<td>430</td>
<td>37</td>
<td>2</td>
</tr>
<tr>
<td>19</td>
<td>289</td>
<td>38</td>
<td>1</td>
</tr>
</tbody>
</table>

As mentioned with the precipitation model, the transitional matrix plays an important role in the Markov chain. Using the wind speed data converted into 38 states, a transitional matrix is a 38 by 38. The transitional matrix is partly shown below.
For the wind speed model, a transition matrix is directly used to decide the wind speed state without calculating a stationary state vector. Also there is no use of a probability distribution to decide wind speed. We decide wind speed using the lower and upper boundaries value for each state.

Once the wind speed state is decided by the Markov chain, wind speed can be calculated by following equation.

\[
V = V_l + Z_i(V_u - V_l)
\]

(5.5)

where \(V_l\) and \(V_u\) are lower and upper boundaries value for each state and \(Z_i\) is the random number from the uniform distribution between 0 and 1. By this equation, we can calculate wind speed into the ranges of each state.
5.2.2 Simulation results of wind speed

Simulation wind speed data for 30 years are generated using a Markov chain and compared with observed data. Statistical values such as mean, standard deviation, minimum and maximum values and percentiles are calculated from observation and simulation wind speed data. These are presented in Table 5.5.

Table 5.5: Summary statistics of observation and simulation wind speed (m/s)

<table>
<thead>
<tr>
<th></th>
<th>Observation</th>
<th>Simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>14.125</td>
<td>14.092</td>
</tr>
<tr>
<td>SD</td>
<td>5.850</td>
<td>6.152</td>
</tr>
<tr>
<td>Min</td>
<td>2.4</td>
<td>2.006</td>
</tr>
<tr>
<td>Max</td>
<td>39.1</td>
<td>39.563</td>
</tr>
<tr>
<td>Percentile</td>
<td>1</td>
<td>5.4</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>8.0</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>9.6</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>12.8</td>
</tr>
<tr>
<td></td>
<td>75</td>
<td>17.6</td>
</tr>
<tr>
<td></td>
<td>90</td>
<td>22.6</td>
</tr>
<tr>
<td></td>
<td>99</td>
<td>31.1</td>
</tr>
</tbody>
</table>

Wind speeds from observation data range from 2.4 to 39.1 with mean around 14.125. The mean wind speed from observation data, 14.125 m/s is very similar to the mean
from simulation data, which is 14.092 m/s. The standard deviation from simulation data is slightly bigger than from observation data. When it comes to percentiles, except the 1st percentile, the other percentiles show very close values between observation and simulation data. It indicates that wind speed can be decided without using a probability distribution when we divide wind speed into many states. To sum up, we can conclude that a Markov chain model is sufficient to reproduce the characteristics of wind speed.

The histogram of simulation wind speed is superimposed with observation data in Figure 5.5. We can see that the frequency of wind speed from simulation data is very similar to observation data. However, wind speeds between 2m/s and 4m/s, which are included in the first and second states are generated more than observed wind speed.

Figure 5.5: Distribution of observation and simulation wind speed (m/s)
In this chapter, we compare calculation methods for evapotranspiration. Since, there are many calculation methods developed for different climatic conditions, using a proper method to calculate evapotranspiration is important to get the correct value for the drought index. Six methods are compared with FAO-56 PM, which is the standard method for evapotranspiration, by calculating $R^2$, RMSE, MAE, MAXE, VE, CORR, EF, D based on regression analysis.
6.1 Evapotranspiration

Calculation methods for evapotranspiration using weather data have been developed under certain weather conditions. Therefore, selecting proper calculation methods provide good results for evapotranspiration values and helps us get accurate drought index values.

6.1.1 Calculation methods for evapotranspiration

The FAO-56 PM method (Allen et al., 1998) is used to calculate the reference evapotranspiration $ET_0$. This method has been recommended as a standard method to calculate reference evapotranspiration using weather data. It has the following form

$$ET_0 = \frac{0.408\Delta (R_n - G) + \gamma \left[ \frac{900}{T + 273} U_2 (e_s - e_a) \right]}{\Delta + \gamma (1 + 0.34U_2)},$$

where $ET_0$ is reference evapotranspiration (mm/day), $R_n$ is net radiation at the crop surface (MJ/m$^2$day), $G$ is soil heat flux density (MJ/m$^2$day), $T_2$ is mean air temperature at height 2m ($^{\circ}$C), $U_2$ is wind speed at height 2m (m/s), $e_s$ is saturation vapor pressure (kPa), $e_a$ is actual vapor pressure (kPa), $\Delta$ is the slope of the vapor pressure curve (kPa/$^{\circ}$C) and $\gamma$ is a psychometric constant (kPa/$^{\circ}$C). In order to use the FAO-56 PM method, we need to find many unfamiliar parameters as we can see in the form. Thus, we need complete and detailed weather data to use FAO-56 PM.
By considering that we were not able to get all detailed weather data, \( ET_0 \) can be estimated by using the simple methods developed by Blaney-Criddle (1988), Hargreaves-Samani (1985), Jensen-Haise (1963), Linacre (Rosenberg et al., 1983; Burman and Pochop, 1994), \( R_n \)-based method (Irmak et al., 2003), Thorthwaite (1948) and Ture (1961). The Blaney-Criddle equation can be written as follows:

\[
ET_0 = P(0.46T_{mean} + 8.13)
\] (6.2)

where \( T_{mean} \) is mean temperature(\(^oC\)) and \( P \) is the percentage of day light hours associated with latitude and longitude.

The Hargreaves equation can be written as follows:

\[
ET_0 = 0.0023 \times R_a(T_{mean} + 17.80)\sqrt{T_{max} - T_{min}}
\] (6.3)

where \( T_{mean}, T_{max} \) and \( T_{min} \) are mean, maximum and minimum temperatures (\(^oC\)) respectively, and \( R_a \) is extraterrestrial radiation (MJ/m\(^2\)day) computed from latitude in radians, sunset hour angle, distance from sun to earth and solar declination.

The method developed by Jensen-Haise for arid and semiarid regions, has the
following equation:

\[
PET = \frac{1}{38 - \left(2 \times \frac{\text{Elevat}}{305}\right) + 7.6 \frac{50}{\left(e_s(T_{\text{max}}) - e_s(T_{\text{min}})\right)}} \\
\times \left[T_{\text{mean}} - \left(-2.5 - 0.14 \left(e_s(T_{\text{max}}) - e_s(T_{\text{min}})\right) - \frac{\text{Elevat}}{550}\right)\right] R_a
\]

(6.4)

where \(\text{Elevat}\) is altitude (m), \(e_s(T_{\text{min}})\) is saturation vapour pressure (kPa) when it is minimum temperature, \(e_s(T_{\text{max}})\) is saturation vapour pressure (kPa) when it is maximum temperature, \(e_a(T_{\text{min}})\) is actual vapour pressure (kPa) when it is minimum temperature and, \(e_a(T_{\text{max}})\) is actual vapour pressure (kPa) when it is maximum temperature. This method is known to overestimate \(ET_0\) in humid areas and underestimate in arid and semi-arid regions.

The Linacre method can be written as follows:

\[
ET_0 = \frac{700 \left(T_{\text{mean}} + 0.0006Z\right)}{100 - L} + 15 \left(T_{\text{mean}} - T_d\right)
\]

(6.5)

where \(Z\) is altitude (m), \(L\) is latitude in degree and \(T_d\) is dew point.

The \(R_n\)-based radiation method developed by Irmak has the equation:

\[
ET_0 = 0.489 + 0.289 R_n + 0.023 T_{\text{mean}}
\]

(6.6)

where \(R_n\) is net radiation (MJ/m\(^2\)day).
Last, the Turc method can be written as follows:

\[
PET = 0.013 \left( \frac{T_{\text{mean}}}{15 + T_{\text{mean}}} \right) (R_s + 50), \quad RH > 50\%
\]

\[
= 0.013 \left( \frac{T_{\text{mean}}}{15 + T_{\text{mean}}} \right) (R_s + 50) \left( 1 + \frac{50 - RH}{70} \right), \quad RH < 50\%
\]

(6.7)

where \(R_s\) is mean solar radiation (W/m\(^2\) day) and \(RH\) is relative humidity (\%).

Each method has different data requirements. And they all have easier calculation equations and fewer data variables compared to the FAO-56 PM method. In case of the Blaney-Criddle method, because we only need temperature we can easily calculate \(ET_0\).

### 6.1.2 Comparison of calculation methods

The six methods are used to estimate \(ET_0\) values. The \(ET_0\) values are compared with the values evaluated from the FAO-56 PM method. The Pearson’s correlation \((R^2)\), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Maximum Absolute Error (MAXE), Volume Error (VE), CORR, Efficiency (EF) and agreement Index (D) are computed by using the following equations.

\[
R^2 = \frac{\sum_{i=1}^{N} (ET_{EQ} - \bar{ET}_{EQ}) (ET_{FAO} - \bar{ET}_{FAO})}{\sqrt{\sum_{i=1}^{N} (ET_{EQ} - \bar{ET}_{EQ})^2 \sum_{i=1}^{N} (ET_{FAO} - \bar{ET}_{FAO})^2}}^2
\]

(6.8)
CHAPTER 6. RECONNAISSANCE DROUGHT INDEX (RDI)

\[ RMSE = \sqrt{\frac{\sum_{i=1}^{N} (ET_{EQ} - ET_{FAO})^2}{N}} \]  \tag{6.9}

\[ MAE = \frac{1}{N} \sum_{i=1}^{N} |ET_{EQ} - ET_{FAO}| \]  \tag{6.10}

\[ MAXE = \text{MAX} (|ET_{EQ} - ET_{FAO}|)_{i=1}^{N} \]  \tag{6.11}

\[ VE(\%) = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{ET_{EQ} - ET_{FAO}}{ET_{EQ}} \right| \]  \tag{6.12}

\[ CORR(\%) = \frac{COV(ET_{EQ}, ET_{FAO})}{\sigma_{ET_{EQ}} \times \sigma_{ET_{FAO}}} \]  \tag{6.13}

\[ EF = \frac{\sum_{i=1}^{N} (ET_{EQ} - \bar{ET}_{EQ})^2 - \sum_{i=1}^{N} (ET_{EQ} - ET_{FAO})^2}{\sum_{i=1}^{N} (ET_{EQ} - \bar{ET}_{EQ})^2} \]  \tag{6.14}

\[ D = 1 - \frac{\sum_{i=1}^{N} (|ET_{FAO} - \bar{ET}_{EQ}| + |ET_{EQ} - \bar{ET}_{EQ}|)^2}{\sum_{i=1}^{N} (ET_{FAO} - ET_{EQ})^2} \]  \tag{6.15}

The smallest value of RMSE, MAE and MAXE indicates a small difference with the FAO-56 PM method. The index of agreement \( D \), developed by Willmott (1981) has a range from 0 to 1 and the value of 1 indicates a perfect match and 0 indicates no match at all.

Furthermore, we apply a linear equation to the six methods and the FAO-56
method. The following linear equation is used.

\[ ET_{EQ} = A + B \times ET_{FAO} \] (6.16)

\( ET_{EQ} \) represents the value of \( ET_0 \) obtained from each simple method and \( ET_{FAO} \) is the value of \( ET_0 \) from the FAO-56 PM method. \( A \) represents the intercept of a linear line and \( B \) represents the slope of a linear line. As a result of regression analysis, we can get the coefficient of determination \( (R^2) \).

### 6.1.3 Results of evapotranspiration

Using the FAO-56 PM method and the six different methods, we calculate daily \( ET_0 \) from 1985 to 2014. Figure 6.1 shows daily \( ET_0 \) values for five years. As with temperature, \( ET_0 \) also shows strong seasonal patterns that repeat every year.
CHAPTER 6. RECONNAISSANCE DROUGHT INDEX (RDI)

(b) Blaney-Criddle

(c) Hargreves-Samani

(d) Jensen-Haise
Figure 6.1: Plot of $ET_0$ in Doha (mm/day)
Figure 6.2 is a boxplot of $ET_0$. The values of $ET_0$ evaluated by FAO-56 PM vary from 0.9811 to 9.7192. Most of the methods, except the Linacre method, look like that they provide similar $ET_0$ values with the FAO-56 PM method. The Linacre method includes the highest value of $ET_0$ and the largest variation from 3.281 to 28.442. So we can say that the Linacre method tends to overestimate $ET_0$ in arid regions. Inversely, the $R_n$-based Radiation method has the smallest range of $ET_0$ values from 1.850 to 5.665. As we mentioned before, we can see that the Jensen-Haise method underestimates $ET_0$ values in arid regions.

Figure 6.2: Comparative boxplots for $ET_0$ calculation methods

Statistical values of $ET_0$ from FAO-56 PM and six methods are summarized in Table 6.1. The minimum, maximum, 25th percentile (Q1), median (Q2), 75th percentile (Q3), maximum, mean and standard deviation are calculated. The largest mean value is provided from the Linacre method and the Jensen-Haise method has
the smallest mean value.

Table 6.1: Summary statistics for $ET_0$ in Doha

<table>
<thead>
<tr>
<th></th>
<th>PM</th>
<th>BC</th>
<th>Har</th>
<th>JH</th>
<th>L</th>
<th>$R_n$</th>
<th>Turc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>0.9811</td>
<td>2.858</td>
<td>0.5479</td>
<td>0.0741</td>
<td>3.281</td>
<td>1.850</td>
<td>0.9128</td>
</tr>
<tr>
<td>Q1</td>
<td>3.0215</td>
<td>5.469</td>
<td>2.2547</td>
<td>1.0589</td>
<td>8.500</td>
<td>2.804</td>
<td>2.4083</td>
</tr>
<tr>
<td>Q3</td>
<td>5.5587</td>
<td>8.175</td>
<td>4.7018</td>
<td>4.1691</td>
<td>17.750</td>
<td>4.377</td>
<td>4.1193</td>
</tr>
<tr>
<td>SD</td>
<td>1.6156</td>
<td>1.5173</td>
<td>1.3886</td>
<td>1.8628</td>
<td>5.1901</td>
<td>0.8678</td>
<td>1.0013</td>
</tr>
</tbody>
</table>

The Pearson’s correlation ($R^2$), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Maximum Absolute Error (MAXE), Volume Error (VE), CORR, Efficiency (EF) and Agreement Index (D) are presented in Table 6.2. The Pearson coefficient varies from 0.4367 to 0.7623. The Turc method has the highest $R^2=0.7623$, while the Linacre method has the lowest $R^2=0.4367$ with the FAO-56 PM method. The smallest RMSE, 1.1590, is obtained from the $R_n$-based radiation method and other methods have relatively small values of RMSE, 2.8507, 1.6317, 2.4806 and 1.4817, for Blaney-Criddel, Hargreaves-Samani, Jensen-Haise and Turc methods, respectively. However, RMSE from the Linacre method is quite large, 10.3425. According to $R^2$ and RMSE, $R_n$-based radiation and Turc method are selected by the method that can be used instead of FAO-56 PM.
Table 6.2: Statistics for the comparison of $ET_0$ methods

<table>
<thead>
<tr>
<th></th>
<th>BC</th>
<th>Har</th>
<th>JH</th>
<th>L</th>
<th>Rn</th>
<th>Turc</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>0.5107</td>
<td>0.7283</td>
<td>0.5765</td>
<td>0.4367</td>
<td>0.7133</td>
<td>0.7623</td>
</tr>
<tr>
<td>RMSE</td>
<td>2.8507</td>
<td>1.6317</td>
<td>2.4806</td>
<td>10.3425</td>
<td>1.159</td>
<td>1.4817</td>
</tr>
<tr>
<td>MAE</td>
<td>2.4389</td>
<td>1.3574</td>
<td>2.1244</td>
<td>8.9496</td>
<td>0.924</td>
<td>1.2232</td>
</tr>
<tr>
<td>MAXE</td>
<td>5.636</td>
<td>3.8584</td>
<td>5.8988</td>
<td>24.0355</td>
<td>2.5564</td>
<td>3.4935</td>
</tr>
<tr>
<td>VE(%)</td>
<td>32.4515</td>
<td>58.0041</td>
<td>212.7213</td>
<td>60.6685</td>
<td>32.0721</td>
<td>49.865</td>
</tr>
<tr>
<td>CORR(%)</td>
<td>71.6085</td>
<td>85.5113</td>
<td>76.081</td>
<td>66.2179</td>
<td>84.6315</td>
<td>87.4888</td>
</tr>
<tr>
<td>EF</td>
<td>-201209</td>
<td>0.2542</td>
<td>-0.2059</td>
<td>-2.6341</td>
<td>-1.0958</td>
<td>-0.967</td>
</tr>
<tr>
<td>D</td>
<td>0.5945</td>
<td>0.8448</td>
<td>0.7164</td>
<td>0.4892</td>
<td>0.7525</td>
<td>0.7465</td>
</tr>
</tbody>
</table>

The scatter plots with linear equations from regression analysis and $R^2$ are presented in Figure 6.4. $ET_0$ values from six methods are plotted against those from the FAO-56 PM method. Turc method has a good agreement (slope=0.542, $R^2=0.7623$) with the FAO-56 PM method. Thus, the importance of temperature, solar radiation and relative humidity is clear for the high correlation of $ET_0$ values. The variables in the Turc method are easily obtainable over arid region such as Doha, making the method suitable for common use.
CHAPTER 6. RECONNAISSANCE DROUGHT INDEX (RDI)

(a) Blaney-criddle

(b) Hargreves-Samani

(c) Jensen-Haise


(a) Linacre

(b) $R_n$-based Radiation

(c) Turc

Figure 6.4: Scatter plots of $ET_0$
6.1.4 Comparison of evapotranspiration methods for RDI

As the results of comparison evapotranspiration methods, Turc method shows the best results in $R^2$ and $CORR$, but we can see comparable results from other methods. So we compare RDI values using four evapotranspiration methods, PM, Hargreaves, $R_n$ and Turc method to see how different they are in calculating of RDI.

Figure 6.5 shows RDI values using different evapotranspiration methods. Compared to PM method, three simple calculation methods provide overestimated values of RDI. Turc method chosen as an alternative method shows the biggest difference in RDI values and Hargreaves method gives the smallest difference in RDI. Even though Hargreaves method provide small differences to PM method in calculating RDI values, we decide to use Turc method to use simulation of humidity.

Figure 6.5: Comparison of RDI in Doha from different methods of $ET_0$


### 6.2 Crop evapotranspiration

#### 6.2.1 Crop coefficient

The crop coefficient is different for each crop type and changes over the growing period. FAO provides general planting and harvest date and crop coefficients for Qatar. In Table 6.3, the growing period and crop coefficients are provided. The growing period is separated into four stages; initial, development, mid and late season. Typically, crop coefficients in mid-season have the largest values.

Most crops in Qatar are grown in winter months and harvested before the start of the summer months, because in summer there are very few rainy days.

Table 6.3: Growing periods and crop coefficients in Qatar (FAO 56, 1988)

<table>
<thead>
<tr>
<th>Crop</th>
<th>Growing period</th>
<th>Growth stage</th>
<th>$k_c$ (Crop coefficient)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Planting</td>
<td>Harvest</td>
<td>Initial</td>
</tr>
<tr>
<td>Alfalfa</td>
<td>1-Jan</td>
<td>2-Mar</td>
<td>10</td>
</tr>
<tr>
<td>Bean</td>
<td>15-Sep</td>
<td>29-Nov</td>
<td>15</td>
</tr>
<tr>
<td>Carrot</td>
<td>15-Oct</td>
<td>12-Feb</td>
<td>20</td>
</tr>
<tr>
<td>Cucumber</td>
<td>15-Nov</td>
<td>25-Mar</td>
<td>25</td>
</tr>
<tr>
<td>Maize</td>
<td>1-Jan</td>
<td>21-May</td>
<td>25</td>
</tr>
<tr>
<td>Onion</td>
<td>15-Oct</td>
<td>31-May</td>
<td>20</td>
</tr>
<tr>
<td>Potato</td>
<td>1-Dec</td>
<td>20-Apr</td>
<td>30</td>
</tr>
<tr>
<td>Rice</td>
<td>1-Dec</td>
<td>30-Apr</td>
<td>30</td>
</tr>
<tr>
<td>Tomato</td>
<td>1-Jan</td>
<td>16-May</td>
<td>30</td>
</tr>
</tbody>
</table>
Table 6.3: Growing periods and crop coefficients in Qatar (FAO 56, 1988)

<table>
<thead>
<tr>
<th>Crop</th>
<th>Growing period</th>
<th>Growth stage</th>
<th>$k_c$ (Crop coefficient)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Planting</td>
<td>Harvest</td>
<td>Initial</td>
</tr>
<tr>
<td>Wheat</td>
<td>15-Dec</td>
<td>24-May</td>
<td>20</td>
</tr>
</tbody>
</table>

6.2.2 Results of crop evapotranspiration

In the previous section where we compared calculation methods for evapotranspiration, the Turc method was chosen as an alternative calculation method for evapotranspiration, it is used to calculate $ET_0$. Once we calculate $ET_0$ using the Turc method, we can calculate the crop evapotranspiration $ET_c$ value by multiplying $ET_0$ by the crop coefficient. The crop coefficient, $k_c$, will be used to decide different values of evapotranspiration depending on crops and growth stages. In this study, crop evapotranspiration for six crops, alfalfa, bean, carrot, maize, tomato and wheat are calculated.

Crop evapotranspiration for six crops are calculated using observed and simulated weather data. Observed weather data from 1985 to 2013 and simulation data with the same length as observation data is used. Mean, standard deviation, minimum and maximum of crop evapotranspiration are presented in Table 6.4. The mean of crop evapotranspiration from tomato shows the largest value and alfalfa has the smallest
It is easy to compare the distribution of crop evapotranspiration between observation and simulation by plotting a boxplot. Crop evapotranspiration calculated from simulation data tends to be smaller than from observation.

<table>
<thead>
<tr>
<th>Crop</th>
<th>Observation</th>
<th>Simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Alfalfa</td>
<td>1.6790</td>
<td>0.6127</td>
</tr>
<tr>
<td>Bean</td>
<td>2.6640</td>
<td>0.6936</td>
</tr>
<tr>
<td>Carrot</td>
<td>2.1840</td>
<td>0.3488</td>
</tr>
<tr>
<td>Maize</td>
<td>2.7510</td>
<td>1.2636</td>
</tr>
<tr>
<td>Tomato</td>
<td>2.8830</td>
<td>1.0954</td>
</tr>
<tr>
<td>Wheat</td>
<td>2.5730</td>
<td>1.1111</td>
</tr>
</tbody>
</table>
We can compare crop evapotranspiration by growth stages. Since crop coefficients have different values depending on the growth stage, it is reasonable that crop evapotranspiration is different by growth stage. The Mean values of crop evapotranspiration by growth stage are summarized in Table 6.5. Most of crops except tomato have the largest value in mid-season.
Table 6.5: Mean of $ET_c$ by growing stage

<table>
<thead>
<tr>
<th>Crop</th>
<th>Observed</th>
<th>Simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Initial Mid Late</td>
<td>Initial Mid Late</td>
</tr>
<tr>
<td>Alfalfa</td>
<td>0.8019 1.9860 1.6616</td>
<td>0.5370 1.4128 1.1892</td>
</tr>
<tr>
<td>Bean</td>
<td>1.8039 3.2297 2.3875</td>
<td>1.4883 2.4692 1.7785</td>
</tr>
<tr>
<td>Carrot</td>
<td>2.0263 2.3110 2.0702</td>
<td>1.5553 1.6166 1.4660</td>
</tr>
<tr>
<td>Maize</td>
<td>0.5988 3.1438 3.2989</td>
<td>0.4097 2.4871 3.2156</td>
</tr>
<tr>
<td>Tomato</td>
<td>1.2107 3.0125 3.6744</td>
<td>0.8247 2.3854 3.5550</td>
</tr>
<tr>
<td>Wheat</td>
<td>1.4138 2.9007 2.4532</td>
<td>0.9284 2.2758 2.4151</td>
</tr>
</tbody>
</table>

Boxplots in Figure 6.7 compare crop evapotranspiration between observation and simulation by growth stage. Boxplots with a white color indicates observed data and boxplots colored gray indicates simulation data. We can easily see that crop evapotranspiration calculated from simulation data is smaller compared to observation.
CHAPTER 6. RECONNAISSANCE DROUGHT INDEX (RDI)
6.3 Calculation of RDI for crops

RDI, which is a measurement of drought, can be used to develop weather derivatives related to drought. To calculate adjusted RDI, we need precipitation and actual evapotranspiration. We calculate adjusted RDI using observation data and simulation data then compare these values.

All RDI values are presented in Appendix C. Table 6.6 shows summary statistics.
of RDI from observation and simulation data. Mean, standard deviation, minimum and maximum are presented.

| Crop    | Observation | | | | Simulation |
|---------|-------------|---|---|---|---|---|---|
|         | Mean  | SD   | Min  | Max  | Mean  | SD   | Min  | Max  |
| Alfalfa | 0.1307 | 0.2805 | 0.0000 | 1.3680 | 0.2349 | 0.4103 | 0.0000 | 1.8390 |
| Bean    | 0.0520 | 0.1851 | 0.0000 | 0.9926 | 0.0471 | 0.1798 | 0.0000 | 0.9482 |
| Carrot  | 0.1268 | 0.1827 | 0.0000 | 0.7161 | 0.1460 | 0.2133 | 0.0000 | 0.7788 |
| Maize   | 0.0700 | 0.1104 | 0.0000 | 0.4678 | 0.1141 | 0.1827 | 0.0000 | 0.7602 |
| Tomato  | 0.0692 | 0.1093 | 0.0000 | 0.4631 | 0.1130 | 0.1813 | 0.0000 | 0.7567 |
| Wheat   | 0.0799 | 0.1047 | 0.0000 | 0.4303 | 0.1153 | 0.1823 | 0.0000 | 0.7341 |

From Table 6.6, we can see all crops in Qatar are grown under the dry condition because the mean of RDI for each crop calculated from observation is very small, close to 0. It means that there is a severe shortage of water from precipitation to grow crops. Since there is no rainy day during the growing period, the minimum value of RDI for all crops is 0.

As comparing summary statistics of RDI from observation and simulation, we can see that RDI from simulation is slightly greater than from observation. It means that dryness contained in simulation data is less severe than historical weather data.
Chapter 7

Conclusion

In this study, we propose a stochastic model for daily temperature, precipitation, humidity and wind speed for Qatar. And RDI for six crops is calculated for the use of pricing weather derivatives related to drought.

7.1 Conclusion

Based on daily modeling for weather derivatives, we examined stochastic weather models for daily temperature, precipitation, humidity and wind speed. Furthermore, calculation methods for evapotranspiration, which is an important factor to calculate RDI, are evaluated to get accurate values for evapotranspiration.

Daily temperature is modeled by an adjusted O-U process, as a representative of mean-reversion process. Since temperature has strong seasonality, a mean-reversion
process is a good choice to model daily temperature. The parameters in an adjusted
O-U process are estimated by the least squares method and maximum likelihood
estimation. By comparing the statistical values such as mean, standard deviation
and percentiles, the adjusted O-U process shows good promise to provide simulation
data that contains the similar properties of temperature.

For precipitation models, a Markov chain is used for precipitation occurrence and a
probability distribution is used to decide the precipitation amount on wet days. More
specifically, a first-order, two-state Markov chain is applied to produce precipitation
occurrence. And the Exponential distribution, Gamma distribution, Lognormal dis-
tribution, GEVD and K4D are tested to fit the precipitation amount. The GEVD
and K4D are selected to decide the precipitation amount on wet days. Especially,
these two distributions show good performance to reproduce extreme amount of pre-
cipitation.

The same approach to temperature is applied for a humidity model. It is reason-
able because humidity also shows the same seasonality and trends as temperature, but
these are less than temperature. As we can expect, an adjusted O-U process works
well to model humidity. However, if we wish to consider the relationship between
temperature and humidity, a new model may be needed.

For daily wind speed, a first-order Markov chain is used as a model. For a wind
speed model, according to minimum and maximum values, 38 states are used in a
Markov chain. Our wind speed model with a Markov chain shows good results except
the case of very small wind speed.

Before calculating the drought index, RDI, a comparison of calculation methods for evapotranspiration was done. To get accurate values for evapotranspiration, six different calculation methods are compared to the FAO-56 PM method. They are compared using $R^2$, RMSE, MAE, MAXE, VE, CORR, EF, D based on regression analysis. Among the six methods, Turc and $R_n$-base radiation methods are selected as an alternative method to FAO-56 PM. The Turc method provides the highest value of $R^2$, 0.7623 and the smallest value of RMSE, MAE and MAXE is provided from $R_n$-based radiation method. Consequently, it is apparent that temperature, solar radiation and relative humidity are highly related to decide evapotranspiration in Qatar, as seen in the two calculation methods. Because both methods have less data requirements than the FAO-56 PM method and the parameters in two methods are easily found from weather data, they are more useful than FAO-56 PM method. We provide improved crop evapotranspiration value using different crop coefficient according to the growth stages.

Finally, RDI for six crops is calculated using observation and simulated weather data. RDI from observation is small enough to categorize Qatar as an arid region. As expected, weather conditions in Qatar are quite dry so it is essential to get help from irrigation systems to grow crops. When we compared RDI from observation and simulation data, RDI from simulation is slightly greater than RDI from observation. It means that the dryness involved in simulation data is less severe than in observa-
tion. These results can be expected also when we calculate crop evapotranspiration because crop evapotranspiration calculated from simulation data is smaller than from observations.

In this study, we only focus on providing a stochastic weather model and drought index values that can be used for weather derivatives. The stochastic weather model we propose seems to work well when we evaluate each model independently. A very close value of the mean is produced from all simulation data but all models underestimate the standard deviation. This is a common pitfall of daily models often reported in literature (Musshoff et al., 2006; Dubrovsky et al., 2004). When we calculate RDI by combining simulation data, it turns out that simulation data shows less of the dryness weather condition compared to observation.

7.1.1 Future work

In daily modelling, a proper weather model that can provide good simulation data is very important. Because simulation weather data is used to calculate index values and this index value is used to calculate the price of weather derivatives, poor weather models can lead to mispricing of weather derivatives.

Increasing the amount of data should improve model performance by giving more information on weather. At the same time, development of parameter estimation methods also leads to improved performance of the model.
Through preprocessing process to replace missing and unreasonable values, we can have good quality data. Especially for weather data, most of weather factors affect to each other, considering the relationship to other weather factors can be a good way to replace missing values.

For humidity model, a mean reversion process is applied to model humidity and it showed good results, with the exception that it provides values that are out of range. When we compare temperature and humidity, they show inverse movements to each other. Generally, when temperature is high humidity is likely to be low. Therefore, a model incorporating in temperature can improve humidity model performance.

For future work, we can calculate the price of weather derivatives for drought using RDI. As we calculate RDI for each crop, we are able to find the price of weather derivatives for each crop. Then we can provide possible outcomes for the price of weather derivatives related to drought for different crops.
Appendix A

References


REFERENCES


REFERENCES


Appendix B

Itô’s Lemma

Using the Itô integral, we can state Itô’s Lemma, which provides a smooth function for dealing with an SDE. Itô’s Lemma is defined as follows. Suppose a stochastic variable $X_t$ satisfies the following stochastic differential equation

$$dX_t = f(X_t, t)dt + g(X_t, t)dW_t$$

(B.1)

and the given function $F = F(X_t, t)$ is a function with continuous partial derivatives $\frac{\partial F}{\partial t}, \frac{\partial F}{\partial X_t}$ and $\frac{\partial^2 F}{\partial X_t^2}$. Then by Itô’s Lemma, $F = F(X_t, t)$ satisfies

$$dF = \frac{\partial F}{\partial t} dt + \frac{\partial F}{\partial X_t} dX_t + \frac{1}{2} g^2(X_t, t) \frac{\partial^2 F}{\partial X_t^2} dt$$

(B.2)

$$= \left( \frac{\partial F}{\partial t} + \frac{1}{2} g^2(X_t, t) \frac{\partial^2 F}{\partial X_t^2} \right) dt + \frac{\partial F}{\partial X_t} dX_t$$
The following is Itô’s isometry that enables us to calculate the expected value and variance for stochastic process.

\[ E \left( \int_a^b \phi(t, \omega) dW_t \right) = 0, \quad (B.3) \]

\[ E \left[ \left( \int_a^b \phi(t, \omega) dW_t \right)^2 \right] = E \left[ \int_a^b \phi^2(t, \omega) dW_t \right]. \quad (B.4) \]
Appendix C

RDI values

<table>
<thead>
<tr>
<th>year</th>
<th>Alfalfa</th>
<th>Bean</th>
<th>Carrot</th>
<th>Maize</th>
<th>Tomato</th>
<th>Wheat</th>
</tr>
</thead>
<tbody>
<tr>
<td>1985</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>1986</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>1987</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>1988</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>1989</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0975</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0617</td>
</tr>
<tr>
<td>1990</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>1991</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>1992</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>1993</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>1994</td>
<td>0.0000</td>
<td>0.9926</td>
<td>0.7161</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.4303</td>
</tr>
<tr>
<td>1995</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.5387</td>
<td>0.4678</td>
<td>0.4631</td>
<td>0.3470</td>
</tr>
<tr>
<td>1996</td>
<td>1.3680</td>
<td>0.0000</td>
<td>0.3657</td>
<td>0.3621</td>
<td>0.1545</td>
<td>0.1432</td>
</tr>
<tr>
<td>1997</td>
<td>0.0000</td>
<td>0.1428</td>
<td>0.1108</td>
<td>0.1567</td>
<td>0.1545</td>
<td>0.0000</td>
</tr>
<tr>
<td>Year</td>
<td>Alfalfa</td>
<td>Bean</td>
<td>Carrot</td>
<td>Maize</td>
<td>Tomato</td>
<td>Wheat</td>
</tr>
<tr>
<td>------</td>
<td>---------</td>
<td>------</td>
<td>--------</td>
<td>-------</td>
<td>--------</td>
<td>-------</td>
</tr>
<tr>
<td>1998</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>1999</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0057</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0036</td>
</tr>
<tr>
<td>2000</td>
<td>0.0075</td>
<td>0.0794</td>
<td>0.2516</td>
<td>0.0019</td>
<td>0.0019</td>
<td>0.1209</td>
</tr>
<tr>
<td>2001</td>
<td>0.0323</td>
<td>0.0000</td>
<td>0.1016</td>
<td>0.0322</td>
<td>0.0318</td>
<td>0.1167</td>
</tr>
<tr>
<td>2002</td>
<td>0.3064</td>
<td>0.0000</td>
<td>0.0290</td>
<td>0.1240</td>
<td>0.1233</td>
<td>0.1412</td>
</tr>
<tr>
<td>2003</td>
<td>0.2707</td>
<td>0.0000</td>
<td>0.1542</td>
<td>0.1357</td>
<td>0.1344</td>
<td>0.0422</td>
</tr>
<tr>
<td>2004</td>
<td>0.1515</td>
<td>0.0432</td>
<td>0.2625</td>
<td>0.0451</td>
<td>0.0447</td>
<td>0.1382</td>
</tr>
<tr>
<td>2005</td>
<td>0.4562</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.1142</td>
<td>0.1130</td>
<td>0.0648</td>
</tr>
<tr>
<td>2006</td>
<td>0.2094</td>
<td>0.0302</td>
<td>0.4227</td>
<td>0.0698</td>
<td>0.0692</td>
<td>0.1323</td>
</tr>
<tr>
<td>2007</td>
<td>0.2182</td>
<td>0.0000</td>
<td>0.1749</td>
<td>0.0684</td>
<td>0.0677</td>
<td>0.1068</td>
</tr>
<tr>
<td>2008</td>
<td>0.4823</td>
<td>0.0000</td>
<td>0.0355</td>
<td>0.1122</td>
<td>0.1114</td>
<td>0.0699</td>
</tr>
<tr>
<td>2009</td>
<td>0.0497</td>
<td>0.0000</td>
<td>0.1725</td>
<td>0.0750</td>
<td>0.0741</td>
<td>0.0866</td>
</tr>
<tr>
<td>2010</td>
<td>0.0533</td>
<td>0.0158</td>
<td>0.0319</td>
<td>0.0709</td>
<td>0.0684</td>
<td>0.0279</td>
</tr>
<tr>
<td>2011</td>
<td>0.0544</td>
<td>0.0699</td>
<td>0.1224</td>
<td>0.0300</td>
<td>0.0296</td>
<td>0.0147</td>
</tr>
<tr>
<td>2012</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0704</td>
<td>0.0156</td>
<td>0.0154</td>
<td>0.1103</td>
</tr>
<tr>
<td>2013</td>
<td>0.0000</td>
<td>0.1364</td>
<td>0.0747</td>
<td>0.0742</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Mean  | 0.1262 | 0.0521 | 0.1178 | 0.0676 | 0.0668 | 0.0771 |
| SD    | 0.2765 | 0.1852 | 0.1790 | 0.1092 | 0.1081 | 0.1039 |
| Max   | 1.3680 | 0.9926 | 0.7161 | 0.4678 | 0.4631 | 0.4303 |

Because carrot and tomato are grown over a period of two years, the number of RDI value for these crops is one less than the others.
Appendix D

R code

#####################################################
######## Temperature model#####################################################
######## Parameter estimation in the SDE
######## Least square method

least_temp <- NA

for (i in 1:29)
{

year <- test[test$year==1982+i,]

est_least <- matrix(NA, nc=4, nr=12)

for (j in 1:12)
{


hh <- year[year$month==j,]$meant

hh_1 <- hh[-1]

hh <- hh[-(length(hh))]

reg <- lm(hh~hh)_1

n <- length(hh)

a <- reg$coef[1]

b <- reg$coef[2]

sd_e <- sqrt(deviance(reg)/df.residual(reg))

dt <- 1

lambda <- -log(a)/dt

mu <- b/(1-a)

sigma <- sd_e*sqrt(-2*log(a)/(dt*(1-a^2)))

est_least[j,1] <-j

est_least[j,2] <-lambda

est_least[j,3] <-mu

est_least[j,4] <-sigma

}
temp_1 <- \texttt{cbind}(1982+i, est\_least)

least\_temp <- \texttt{rbind}(least\_temp, temp_1)

\texttt{colnames}(least\_temp) <- \texttt{c(}"year","month","lambda","mu","sigma"

least\_temp <- (least\_temp[-1,])

\begin{verbatim}

####### MLE method

mle\_temp <- \texttt{NA}

\texttt{for} ( i \texttt{in} 1:29 )
{

    year <- test [ test\$year==1982+i ,]

    est\_mle <- \texttt{matrix}(\texttt{NA}, nc=4, nr=12)

    \texttt{for} ( j \texttt{in} 1:12 )
{

        hh <- year [year\$month==j ,]\$meant

        hh\_1 <- hh[-1]

        hh <- hh[-(\texttt{length}(hh))]

}\end{verbatim}
R CODE

```r
Sx <- sum(hh)
Sy <- sum(hh_1)
Sx <- as.numeric(Sx)
Sy <- as.numeric(Sy)

Sxx <- sum((hh)^2)
Sxy <- sum((hh_1)*(hh))
Sxy <- as.numeric(Sxy)
Syy <- sum((hh_1)^2)

n <- length(hh)
delta <- 1
mu <- (Sy*Sxx-Sx*Sxy)/(n*(Sxx-Sxy) - ((Sx^2) - Sx*Sy))
lambda <- (-1/delta)*log(((Sxy-mu*Sx-mu*Sy+n*mu^2)/(Sxx-2*mu*Sx+n*mu^2)))
alpha <- exp(-lambda*delta)

sigma_2_hat <- (1/n)*(Syy-2*alpha*Sxy+alpha^2*Sxx-2*
mu*(1-alpha)*(Sy-alpha*Sx)+n*(mu^2)*(1-alpha)^2)

sigma_2 <- 2*lambda/(1-(alpha^2))*sigma_2_hat
```
est.mle[j,1] <- j
est.mle[j,2] <- lambda
est.mle[j,3] <- mu
est.mle[j,4] <- sqrt(sigma_2)
}
temp.1 <- cbind(1982+i, est.mle)

mle.temp <- rbind(mle.temp, temp.1)
}
colnames(mle.temp) <- c("year","month","lambda","mu","sigma")
mle.temp <- (mle.temp[,-1,])

#### OU simulation

#### Parameters

least <- least.temp[-1,]
est <- data.frame(least)
est <- aggregate(cbind(mu, lambda, sigma) ~ month, est, mean)
mle <- mle.temp[-1,]
est <- mle
est <- aggregate(cbind(mu, lambda, sigma) ~ month, est, mean)
tmean <- aggregate(mean ~ month, test, mean)

########### Simulation function base on the solution of SDE

OU <- function(n, mu, lambda, sigma, x0)
{
  dt <- 1
  x <- c(x0)
  for (i in 2:(n+1))
  {
    x[i] <- x[i-1]*exp(-lambda*dt) + mu*(1-exp(-lambda*dt)) +
    sigma*sqrt((1-exp(-2*lambda*dt))/(2*lambda)) *rnorm(1,0,1)
  }
  return(x);
}

########### 365 day data

tem2 <- NA

n <- c(31,28,31,30,31,30,31,31,30,31,30,31)
for (k in 1:29) {

  tem <- NA
  tem1 <- NA

  for (j in 1:12) {
    if (j == 1 | j == 3 | j == 5 | j == 7 | j == 8 | j == 10 | j == 12) {
      n <- 31
    } else if (j == 2) {
      n <- 28
    } else {
      n <- 30
    }
    tem <- OU(n, est$mu[j], est$lambda[j], est$sigma[j],
              tmean[j,2])
    tem1 <- c(tem1, tem[-1])
  }
  tem2 <- c(tem2, tem1[-1])
}

# Precipitation model
# Transitional matrix
occ <- NULL
occ[1] <- 1

# Define state
# 1 <- dry day following a dry day
# 2 <- wet day following a dry day
# 3 <- dry day following a wet day
# 4 <- wet day following a wet day

for (i in 1:nrow(doha_30)) {
  if (doha_30$poi[i]==0) {
    if (doha_30$poi[i+1]==0) {occ[i+1] <- 1}
    else {occ[i+1] <- 2}
  } else if (doha_30$poi[i+1]==0) {occ[i+1] <- 3}
  else {occ[i+1] <- 4}
}

doha_30 <- cbind(doha_30, occ)
### Transitional matrix and stationary vector for each month

# 0 dry, 1 wet

\[
p \leftarrow \text{matrix}(\text{NA}, \text{nc}=7, \text{nr}=12)
\]

\[
\text{colnames}(p) \leftarrow \text{c( "month", "p00", "p01", "p10", "p11", "pd", "pw")}
\]

\[
\text{for ( } i \text{ in } 1:12)
\]

\{

\[
\text{data} \leftarrow \text{doha}_30[\text{doha}_30[,2]==i ,]
\]

\[
\text{fr} \leftarrow \text{table(data$occ)}
\]

\[
p[i,1] \leftarrow i
\]

# Transitional matrix

\[
p[i,2] \leftarrow \text{fr["1"]/(fr["1"]+fr["2"]) \# p00}
\]

\[
p[i,3] \leftarrow \text{fr["2"]/(fr["1"]+fr["2"]) \# p01}
\]

\[
p[i,4] \leftarrow \text{fr["3"]/(fr["3"]+fr["4"]) \# p10}
\]

\[
p[i,5] \leftarrow \text{fr["4"]/(fr["3"]+fr["4"]) \# p11}
\]

# Stationary vector

\[
p[i,6] \leftarrow p[i,4]/(p[i,4]+p[i,3]) \# sd
\]

\[
p[i,7] \leftarrow p[i,3]/(p[i,4]+p[i,3]) \# sw
\]
### Simulation of precipitation occurrence

temp <- c(0,0,0,0)

r <- matrix(NA, nc=12, nr=31)

for (k in 1:31)
{
  for (i in 1:12)
  {
    if (i == 1 | i == 3 | i == 5 | i == 7 | i == 8 | i == 10 | i == 12)
      { nd <- 31 }
    else if (i == 2)
      { nd <- 28 }
    else
      { nd <- 30 }

    for (j in 1:nd)
    {
      r[j, i] <- 999 # set initial value
    }
  }

  colnames(r) <- c(1,2,3,4,5,6,7,8,9,10,11,12)
  n_pp <- melt(r, na.rm=T)
  n_pp <- cbind(k, n_pp)
}

temp <- rbind(temp, n_pp)
}
colnames(temp) <- c("senario", "day", "month", "occ")

### Precipitation occurrence using stationary vectors

for (l in 1:(nrow(temp)-1))
{
  if (runif(1) < p[temp$month[l+1], 6]) { temp[l+1,4] <- 0}
  else { temp[l+1,4] <- 1}
}
s_pp <- temp[-1,]
colnames(s_pp) <- c("senario", "day", "month", "occ")

### Simulation of precipitation amount

s_pp[,5] <- NA
colnames(s_pp)[5] <- c("pp")

for (i in 1:nrow(s_pp))
{
  if (s_pp[i,4]==1)
  {

R CODE

```r
s_pp[i,5] <- n_quakap(runif(1,max=0.98),l_para) # probability distribution
```

```r
else {s_pp[i,5] <- 0}
```

### Calculation of RDI

# Import data

c <- x # Define crop

kc <- period[c, c(14:16)] # Crop coefficient values

k <- Turc # ET0 data

# Define growth period

sm <- period[c,4] # start month

sd <- period[c,5] # start day

em <- period[c,6] # end month

ed <- period[c,7] # end day

# Define growth stages
ini <- period[,9]

develop <- period[,10]

mid <- period[,11]

end <- period[,12]

# Calculation of crop evapotranspiration
k$etini <- kc[,1]*k$ET0

k$etmid <- kc[,2]*k$ET0

k$etend <- kc[,3]*k$ET0

# find first day and last day of growing period

if (sm < em)

  { s <- which(k$month==sm & k$day==sd)
    e <- which(k$month==em & k$day==(ed-1))
  }

else

  { e <- which(k$month==em & k$day==(ed-1))[1]
    s <- which(k$month==sm & k$day==sd)[-length(e)+1]]

  }

grow <- cbind(s,e)

# Calculate RDI for each year
index <- matrix(NA, nr=nrow(grow))

for (i in 1:nrow(grow))
{

cropet <- k[grow[i,1]:grow[i,2],] # select data for each growing period

cropet$et <- NA

# Decide crop evapotranspiration by crop stage

cropet$et[seq(1, length=ini)] <- cropet$etini[seq(1, length=ini)]
cropet$et[seq(ini+1, length=mid+dev)] <- cropet$etmid[seq(ini+1, length=mid+dev)]
cropet$et[seq(ini+dev+mid+1, length=end)] <- cropet$etend[seq(ini+dev+mid+1, length=end)]

index[i,] <- sum(cropet$pp)/sum(cropet$et) # RDI value
}

# # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #