Implementation of a Hybrid Weather Generator and Creating Sets of Synthetic Weather Series Consistent with Seasonal Climate Forecasts in the Southeastern United States

William Joel Forsee
University of Miami, bforsee@rsmas.miami.edu

Follow this and additional works at: https://scholarlyrepository.miami.edu/oa_theses

Recommended Citation
https://scholarlyrepository.miami.edu/oa_theses/215

This Open access is brought to you for free and open access by the Electronic Theses and Dissertations at Scholarly Repository. It has been accepted for inclusion in Open Access Theses by an authorized administrator of Scholarly Repository. For more information, please contact repository.library@miami.edu.
IMPLEMENTATION OF A HYBRID WEATHER GENERATOR AND CREATING SETS OF SYNTHETIC WEATHER SERIES CONSISTENT WITH SEASONAL CLIMATE FORECASTS IN THE SOUTHEASTERN UNITED STATES

By

William Joel Forsee

A THESIS

Submitted to the Faculty
of the University of Miami
in partial fulfillment of the requirements for
the degree of Master of Science

Coral Gables, Florida

May 2008
UNIVERSITY OF MIAMI

A thesis submitted in partial fulfillment of
the requirements for the degree of
Master of Science

IMPLEMENTATION OF A HYBRID WEATHER GENERATOR AND CREATING
SETS OF SYNTHETIC WEATHER SERIES CONSISTENT WITH SEASONAL
CLIMATE FORECASTS IN THE SOUTHEASTERN UNITED STATES

William Joel Forsee

Approved:

________________                      ____________________
Dr. David Letson                  Dr. Terri A. Scandura
Associate Professor of            Dean of the Graduate School
Marine Affairs and Policy

________________                      ____________________
Dr. Guillermo Podestá             Dr. Kenny Broad
Research Professor of             Assistant Professor of
Meteorology and Physical          Marine Affairs and Policy
Oceanography

________________                      ____________________
Dr. Don Olson                    Dr. Don Olson
Professor of                     Professor of
Meteorology and Physical          Meteorology and Physical
Oceanography
Oceanography
FORSEE, WILLIAM JOEL (M.S., Marine Affairs and Policy)  

Implementation of a Hybrid Weather Generator and Creating Sets of Synthetic Weather Series Consistent with Seasonal Climate Forecasts in the Southeastern United States  

(May 2008)

Abstract of a thesis at the University of Miami.

Thesis supervised by Dr. David Letson.

No. of pages in text. (149)

Stochastic weather generators create multiple series of synthetic daily weather (precipitation, maximum temperature, etc.), and ideally these series will have statistical properties similar to those of the input historical data. The synthetic output has many applications and for example, can be used in sectors such as agriculture and hydrology. This work used a “hybrid” weather generator which consists of a parametric Markov chain for generating precipitation occurrence and a nonparametric k-nearest neighbor method for generating values of maximum temperature, minimum temperature, and precipitation. The hybrid weather generator was implemented and validated for use at 11 different locations in the Southeastern United States. A total of 36 graphic diagnostics were used to assess the model’s performance. These diagnostics revealed that the weather generator successfully created synthetic series with most statistical properties of the historical data including extreme wet and dry spell lengths and days of first and last freeze.

Climate forecasts are typically provided for seasons or months. Alternatively, process models used for risk assessment often operate at daily time scales. If climate forecasts were incorporated into the daily weather input for process models, stakeholders could then use these models to assess possible impacts on their sector of interest due to
anticipated changes in climate conditions. In this work, an “ad hoc” resampling approach was developed to create sets of daily synthetic weather series consistent with seasonal climate forecasts in the Southeastern United States. In this approach, the output of the hybrid weather generator was resampled based on forecasts in two different formats: the commonly used tercile format and a probability distribution function. This resampling approach successfully created sets of synthetic series which reflected different forecast scenarios (i.e. wetter or drier conditions). Distributions of quarterly total precipitation from the resampled synthetic series were found to be shifted with respect to the corresponding historical distributions, and in some cases, the occurrence and intensity statistics of precipitation in the new weather series had changed with respect to the historical values.
Acknowledgements

First, I would like to thank my MPO advisor, Guillermo Podestá, for all of his guidance and patience. He put a lot of resources into ensuring that my graduate experience would advance not just my knowledge but skills needed to succeed in life. I would like to thank my MAF advisor, David Letson, for his support and for enabling me to conduct my graduate work at RSMAS. I would also like to thank my committee members, Kenny Broad and Don Olson, for their suggestions and assistance as I progressed through my coursework and research.

I am grateful to Balaji Rajagopalan and Somkiat Apipattanavis for supplying the weather generator code and to Jim Brown for implementation of the code and programming assistance. I am deeply grateful to the Southeastern Climate Consortium for financial support and David Zierden of Florida State University, who was instrumental in providing the cooperative data set.

I would like to recognize the RSMAS community, a collection of diverse and wonderful people. Particular thanks to Chidong Zhang, who first brought me to RSMAS and has remained a source of support during my time at RSMAS.

Lastly, and most importantly, I would like to thank my father and sister for their support from far away.
Table of Contents

List of Figures vi
List of Tables xi
Chapter 1: Introduction 1
  1.1 Motivation 1
  1.2 Background - Stochastic Weather Generators 5
  1.3 Objectives 8
  1.4 The Study Area 9
Chapter 2: Data and Methods 13
  2.1 Climate Data 13
  2.2 Weather Generators 18
    2.2.1 Parametric Weather Generators 18
    2.2.2 Non-Parametric Weather Generators 21
  2.3 The Semi-Parametric Weather Generator 24
    2.3.1 Markov Chain Model 24
    2.3.2 Nearest Neighbor Algorithm 27
    2.3.3 Changes in the Calculation of Euclidean Distance 31
    2.3.4 Temperature Bias Corrections 31
  2.4 Conditioning Synthetic Series on Climate Forecasts 35
    2.4.1 Forecasting Climate 36
    2.4.2 Climate Forecast Formats 38
    2.4.3 Conditioning Synthetic Series with Weather Generators 42
    2.4.4 A New Approach for Conditioning Synthetic Series with the Semi-Parametric Weather Generator 44
      2.4.4.1 Resampling Using Format I 47
      2.4.4.2 Resampling Using Format II 50
  2.5 Estimation of Tercile Boundaries 53
    2.5.1 Data Preparation 55
    2.5.2 Empirical Quantile Estimation (EMP) 55
    2.5.3 Estimation by Fitting a Gamma Distribution (GAMMA) 56
    2.5.4 Nonparametric Kernel Density Estimation (KERNEL) 57
    2.5.5 Density Estimation Using Adaptive Splines (SPLINE) 58
Chapter 3: Results

3.1 Synthetic Output from the Weather Generator

3.1.1 Temperature Bias

3.1.2 Assessment of Maximum Temperature and Minimum Temperature Generation

3.1.2.1 Monthly Means of Tmax and Tmin

3.1.2.2 Standard Deviation of Tmax and Tmin

3.1.2.3 Lag-1 Correlation for Tmax and Tmin

3.1.2.4 Skewness of Tmax and Tmin

3.1.2.5 Cross Correlation of Tmax and Tmin

3.1.2.6 Dates of First Freeze and Last Freeze (Indicators of Frost Free Period)

3.1.2.7 Hot and Cold Spells

3.1.3 Assessment of Precipitation Generation

3.1.3.1 Probability of a Wet Day

3.1.3.2 Probability of a Very Wet Day

3.1.3.3 Mean Wet and Dry Spell Lengths

3.1.3.4 Longest Annual Wet and Dry Spell Lengths

3.1.3.5 Daily Precipitation Intensity

3.1.3.6 Monthly Total Precipitation

3.1.3.7 Annual Total Precipitation

3.1.3.8 Standard Deviation of Daily Precipitation

3.1.3.9 Standard Deviation of Monthly Total Precipitation

3.2 Synthetic Series Conditioned on Climate Forecasts

3.2.1 Assessment of Tercile Boundary Estimates

3.2.1.1 Q33 and Q66 Estimates of Temperature and Precipitation

3.2.1.2 Divergence of Estimates by Different Methods of Q33 and Q66

3.2.1.3 Impact of Missing Data upon Estimation of Q33 and Q66

3.2.2 Assessment of Precipitation Values from Conditioned Synthetic Series

3.2.2.1 Median Quarterly Total Precipitation

3.2.2.2 Distributions of Quarterly Total Precipitation

3.2.2.3 Q-Q Plots of Distributions of Quarterly Total Precipitation

3.2.2.4 Probability of a Wet Day

3.2.2.5 Median Daily Precipitation
# List of Figures

2.1 Schematic diagram of methodology ........................................... 14
2.2 Map of stations ........................................................................ 16
2.3 Schematic diagram of the semi-parametric generator by Apipattanavis et al. (2007) ................................................................. 30
2.4 Biases of synthetic daily maximum temperature (Tmax) and minimum temperature (Tmin) on dry days in Atlanta (GA) ................. 33
2.5 Biases of synthetic daily maximum temperature (Tmax) and minimum temperature (Tmin) on wet days in Atlanta (GA) .................. 34
2.6 IRI net assessment forecast ....................................................... 40
2.7 CPC probability of exceedance forecast ..................................... 41
2.8 Schematic diagram of resampling procedure using the tercile forecast format ........................................................................ 49
2.9 Histograms of quarterly total precipitation for historical and resampled data ........................................................................... 50
2.10 Schematic diagram of resampling procedure using the PDF forecast format ............................................................................ 54
3.1 Monthly means of daily Tmax and Tmin of synthetic and historical data at Atlanta (GA) ................................................................. 64
3.2 Standard deviation of daily Tmax and Tmin of synthetic and historical data at Atlanta (GA) ............................................................... 66
3.3 Lag-1 correlation of daily Tmax and Tmin of synthetic and historical data for dry days at Clermont (FL) ................................................. 68
3.4 Lag-1 correlation of daily Tmax and Tmin of synthetic and historical data for wet days at Clermont (FL) 69
3.5 Skewness of daily Tmax and Tmin of synthetic and historical data at Mobile (AL) 72
3.6 Cross correlation of daily Tmax and Tmin of synthetic and historical data on dry days at Atlanta (GA) 74
3.7 Cross correlation of daily Tmax and Tmin of synthetic and historical data on wet days for Atlanta (GA) 75
3.8 Average day of first freeze and day of last freeze of synthetic and historical data at Atlanta (GA) 77
3.9 Total hot and cold spells of synthetic and historical data at Mobile (AL) 79
3.10 Probability of a wet day from synthetic and historical data at Atlanta (GA) 82
3.11 Mean wet and dry spell lengths of synthetic and historical data at Miami (FL) 85
3.12 Maximum annual wet and dry spell lengths of synthetic and historical data at Atlanta (GA) 87
3.13 Quantile-quantile plot of daily precipitation of synthetic and historical data at Miami (FL) 88
3.14 Quantile-quantile plot of daily precipitation of synthetic and historical data at Mobile (AL) 89
3.15 Median monthly total precipitation of synthetic and historical data at Tifton (GA) 91
3.16 Mean annual total precipitation of synthetic and historical data at
Atlanta (GA) 93
3.17 Standard deviation of daily precipitation of historical and synthetic data at
Atlanta (GA) 95
3.18 Standard deviation of monthly total precipitation of historical and
synthetic data at Brooklet (GA) 96
3.19 Contour plots of empirically-derived Q33 estimates for quarterly mean
surface temperature 99
3.20 Contour plots of empirically-derived Q66 estimates for quarterly mean
surface temperature 100
3.21 Contour plots of empirically-derived Q33 estimates for quarterly total
precipitation 101
3.22 Contour plots of empirically-derived Q66 estimates for quarterly total
precipitation 102
3.23 Scatterplots of Q66 estimates for JFM precipitation by all methods for all
stations 104
3.24 Quarterly total precipitation during JFM for Lisbon (FL) and the curves
fitted to the empirical distribution by the three methods 106
3.25 Quarterly total precipitation during JFM for Lisbon (FL) in (a) La Niña
years (b) El Niño years and (c) Neutral years 107
3.26 Empirical (not bootstrapped) estimates of Q33 and Q66 from simulated
incomplete data sets of various sample sizes 109
3.27 Median quarterly total precipitation of conditioned synthetic data for very wet and very dry forecasts using format I and for historical data at Chipley (FL)  

3.28 Median quarterly total precipitation of conditioned synthetic data for very wet and very dry forecasts using format II and for historical data at Chipley (FL)  

3.29 Histograms and density plots of quarterly total precipitation of conditioned synthetic data for very wet and very dry forecasts using format II and for historical data at Chipley (FL) during JFM  

3.30 Histograms and density plots of quarterly total precipitation of conditioned synthetic data for very wet and very dry forecasts using format II and for historical data at Chipley (FL) during JFM  

3.31 Histograms and density plots of quarterly total precipitation of conditioned synthetic data for very wet and very dry forecasts using format I and for historical data at Chipley (FL) during JAS  

3.32 Histograms and density plots of quarterly total precipitation of conditioned synthetic data for very wet and very dry forecasts using format II and for historical data at Chipley (FL) during JAS  

3.33 Quantile-quantile plots of quarterly total precipitation for a very dry forecast using format I at Chipley (FL)  

3.34 Quantile-quantile plots of quarterly total precipitation for a very dry scenario using format II at Chipley (FL)
Quantile-quantile plots of quarterly total precipitation for a very wet scenario using format I at Chipley (FL)

Quantile-quantile plots of quarterly total precipitation for a very wet scenario using format II at Chipley (FL)

Probability of a very wet day of conditioned synthetic data for very wet and very dry forecasts using format I and for historical data at Chipley (FL)

Probability of a very wet day of conditioned synthetic data for very wet and very dry forecasts using format II and for historical data at Chipley (FL)

Median daily precipitation of conditioned synthetic data for very wet and very dry forecasts using format I and for historical data at Chipley (FL)

Median daily precipitation of conditioned synthetic data for very wet and very dry forecasts using format II and for historical data at Chipley (FL)
List of Tables

2.1 Cooperative station data sets
2.2 Matrix of transition probabilities
2.3 Hypothetical forecast tercile probabilities
3.1 Values of 80th percentile of monthly distributions of daily precipitation for all locations
3.2 Mean values of annual historical total precipitation and median values of mean annual synthetic total precipitation at all locations
3.3 Hypothetical forecast tercile probabilities for very wet and very dry scenarios
3.4 Values of Q5, Q50, and Q95 for Figures 3.29 and 3.30
3.5 Values of Q5, Q50, and Q95 for Figures 3.31 and 3.32
Chapter 1
Introduction

1.1 Motivation

Climate variability and climate change can have significant impacts on various sectors of society in many parts of the globe (Ropelewski and Halpert 1987; Schmidt et al. 2000; IPCC 2007a; Katz et al. 2003; Ferro et al. 2005). There is much concern about the impacts of fluctuations in the climate system on social and biophysical systems (Hewiston and Crane 1996; Easterling 1999). For example, the global food supply is vulnerable to climate variability and climate change (Hansen et al. 1998; Jones et al. 2000). Methods of assessing local or regional scale impacts of large scale climate change and variability are needed to quantify the vulnerability of social and biophysical systems to these forces.

A major source of climate variability on seasonal-to-interannual scales in many parts of the world is the El Niño-Southern Oscillation (ENSO) (Trenberth 1997). ENSO is a coupled ocean-atmosphere phenomenon that occurs in the tropical Pacific Ocean. It involves two extreme phases: El Niño years (also referred to as “warm events”, because of increased sea surface temperatures in the tropical Pacific Ocean), and La Niña years (or “cold events”); years which do not fall in the extreme phases are referred to as “neutral”. One region influenced by this phenomenon is the Southeastern United States (SEUS), the area targeted in this study, where seasonal mean temperature and precipitation amounts are associated with the phases of ENSO (e. g. Ropelewski and Halpert 1987; Kiladis and Diaz 1989; Sittel 1994).
Advances in understanding and observations of the oceans and atmosphere have made it possible to predict with imperfect but usable skill ENSO-related sea surface temperature (SST) anomalies months in advance (Goddard et al. 2001). In turn, predicted SSTs and atmospheric general circulation models can be used to forecast total precipitation and mean temperature in some seasons and for some regions of the planet (Mason et al. 1999). Climate forecasts display potential climate anomalies on a spatial scale typical of global circulation models, and on temporal scales of one to three months (Hansen and Indeje 2004). These forecasts are being disseminated by several agencies around the world, including the United States’ Climate Prediction Center (CPC) and the International Research Institute for Climate and Society (IRI).

It is often assumed that climate forecasts will benefit climate-sensitive sectors by allowing stakeholders to mitigate negative consequences of climate variability or, alternatively, capitalize on potentially beneficial effects. Nevertheless, several studies have identified theoretical and practical obstacles to the use of climate information and forecasts (Pulwarty and Redmond 1997; Orlove and Tosteson 1999; Stern and Easterling 1999; Broad and Agrawala 2000; Glantz 2001; Broad et al. 2002; Hartmann et al. 2002; Lemos et al. 2002; Patt and Gwata 2002). Some obstacles include procedural, institutional, and cognitive difficulties in receiving/understanding information, or in the ability and willingness of decision-makers to modify their actions. Other obstacles stem from limitations inherent to the climate system’s complexities: forecasts have coarse spatial and temporal resolution (whereas stakeholders often are more interested in information at the scale of their decisions), not all relevant or desired climate variables can be predicted, the skill of forecasts is not well characterized or understood, and
contradictory predictions may coexist. Effective use of seasonal climate forecasts must be based on understanding these constraints and how to overcome them.

Adaptive responses to climate and other risk factors require salient information to support decisions. For example, agricultural outcomes of decisions are more relevant to stakeholders than raw climate information: a farmer often is more interested in receiving likely distributions of crop yields or economic returns than a seasonal precipitation forecast (Hammer et al. 2001). A greater capacity is needed to convert raw climate information (seasonal forecasts, decadal climate projections) into actionable distributions of sectoral outcomes at regional and local scales for risk assessment and management.

Even though global climate models have steadily increased their spatial resolution, a significant gap remains between those scales at which global models are skillful and the scales often relevant to stakeholders’ decisions. Thus, considerable effort has focused on the development of techniques to bridge this gap: techniques to translate climate information from larger to smaller scales are referred to as “downscaling” (Hewiston and Crane 1996). Different approaches have been developed to downscale climate information (Wilby and Wigley 1997; Wilby et al. 1998; IPCC 2007b). They typically fall into one of two streams: high resolution dynamical downscaling involving either variable resolution global climate models (e.g. Giorgi 1990) or regional (or limited area) climate models (e.g. Mearns et al. 1995; Mearns et al. 1999), and empirical or statistical downscaling (e.g. Zorita and von Storch 1999; Mearns et al. 1999; Huth 1999). Examples of empirical and statistical techniques include weather classification schemes (Mearns et al. 1999) and neural networks (Weichert and Bürger 1998). Several works
have compared the performance of a range of approaches (Wilby and Wigley 1997; Zorita and von Storch 1999; Wilby et al. 1998; Fowler et al. 2007).

Another statistical approach to downscaling involves the use of stochastic weather generators (e.g. Yates et al. 2003; Clark et al. 2004; Apipattanavis et al. 2007). These models create multiple synthetic weather series that have statistical properties similar to those of the input data. Their ability to preserve the statistical properties of historical weather makes weather generators an excellent choice for downscaling (Hansen and Ines 2005).

Outcomes of alternative decisions in weather-sensitive sectors of society can be simulated with process models. For example, crop simulation models (Easterling et al. 1992; Wallis and Griffiths 1995; Jones et al. 2003) can be used in agricultural decision making to assess potential productivity and risk for different climate conditions during the growing season (Hansen and Indeje 2004).

Process models often require daily weather series as input. Historical daily data can be used, but historical records often are short or difficult to obtain. More fundamentally, observed sequences are only one realization of the weather process (Richardson 1981). A thorough assessment of different management decisions should explore a range of possible outcomes where multiple different weather sequences, each with the same statistical properties as historical weather, are used as input into these models. Process models could be driven with weather sequences from numerical ocean-atmosphere models used to simulate climate, but spatial averaging across grid cells can alter the temporal variability of daily weather variables (Mearns et al. 1995). Dynamical models with higher resolution do not produce daily values with realistic temporal
structure. For example, Mearns et al. (1996) found that daily precipitation intensity and occurrence in the Great Plains of the United States was not accurately reproduced due to inaccurate simulation of topography; and noted that biases in precipitation generation could significantly impact crop yield predictions. One means of overcoming these obstacles is the use of stochastic weather generators, which can preserve the location-specific characteristics of daily weather.

1.2 Background - Stochastic Weather Generators

Stochastic weather generators are statistical models that create synthetic (i.e. simulated) series of daily weather from historical data. The statistical properties of synthetic series are intended to be similar to those of observed historical weather. Examples include the work of Richardson (1981), Semenov and Porter (1995), Semenov and Barrow (1997), Rajagopalan and Lall (1999), Yates et al. (2003), Schoof et al. (2005), Sharif and Burn (2006), and Schoof (2008). Reviews of commonly used weather generators can be found in Wallis and Griffiths (1995), Johnson et al. (1996), and Semenov et al. (1998).

Common uses of weather generators include filling in missing weather data, crop simulation, watershed planning, simultaneous modeling of weather at multiple sites, and downscaling of climate change projections. In some cases, creation of synthetic weather is done separately from modeling (crop simulation, etc.) and in other cases weather generators are incorporated within different process models. Mearns et al. (1999) performed sensitivity analysis of crop models to different climate scenarios. Multisite simulation of temperature, precipitation, and solar radiation was done by Wilks (1998). Downscaling of climate change projections for watershed analysis was done by Sharif
and Burn (2006). Yates et al. (2003) used weather generators to simulate potential climate scenarios at various locations in the United States.

Several approaches have been proposed for the stochastic generation of weather variables. These approaches can be grouped into two main categories: parametric and nonparametric methods (Sharif and Burn 2006). Parametric weather generators involve assumptions about the statistical properties of the data. In most available parametric models, the generation involves a sequence of three steps: (1) generation of precipitation occurrence (i.e., generating a sequence of “dry” or “wet” days), (2) generation of precipitation intensity (i.e., the rainfall amount on a rainy day), and (3) generation of all other weather variables. Typically, Markov chains have been used to model precipitation occurrence in parametric generators. Other weather variables typically are generated using autoregressive models dependent upon precipitation.

Some disadvantages of using parametric approaches include: (1) the need for prior assumptions about the distributions of the historical data, (2) a large number of parameters must be fitted for each season (and this increases exponentially if simulations are to be conditioned on large scale climate indices), and (3) only linear relationships between the variables can be reproduced (Rajagopalan and Lall 1999; Yates et al. 2003).

An alternative to parametric approaches is the use of non-parametric methods, which are data-driven and do not require assumptions about the distributions of the variables of interest. They can provide a flexible framework, are parsimonious, and can be easily modified to do simulations based on particular climate states (Rajagopalan and Lall 1999; Yates et al. 2003; Clark et al. 2004).
This work will partially focus on a nonparametric k-nearest neighbor algorithm (k-NN) based on the model by Rajagopalan and Lall (1999). Briefly, in the k-NN approach, a scalar distance measurement is used to determine the \( k \) days in the historical data most similar (in multivariate weather space) to a given synthetic day. One of these \( k \) neighbors is selected by resampling, and the values of the variables of the following day to the selected neighbor become the daily values for the following day in the synthetic series. Some of the advantages of using the k-NN method include: (1) the dependence between all variables of interest can be well simulated, (2) lag-one correlations are often captured better than by other methods, and (3) these methods can be easily extended to multisite simulation (Yates et al. 2003).

Weather generators which utilize nearest neighbor resampling sometimes do not accurately reproduce wet and dry spell statistics (Young 1994; Apipattanavis et al. 2007). A possible solution to this problem is to use parametric Markov models to generate precipitation occurrence. Markov models determine the precipitation state for a given day based on probabilities of precipitation occurrence on the previous day(s). Richardson (1981) found that two-state, first-order Markov models were successful in reproducing precipitation occurrence in many cases. In situations where first-order models have not worked, higher-order Markov chains (e.g. Jones and Thornton 1997; Wilks 1999) have been used with success (Katz and Parlange 1998). However, higher-order models require the estimation of a large number of parameters (Hutchinson 1987).

An alternative to increasing the order of Markov models could be to increase the number of precipitation states. Gregory et al. (1992) hypothesized that a first-order, many-state model could be a viable alternative to a two-state, higher-order Markov chain.
They developed a first-order, multi-state Markov chain that simulated the variance of long term precipitation totals and the serial correlation of daily precipitation amounts better than two-state Markov models of order one and two (Gregory et al. 1993).

1.3 Objectives

The overarching goal of this work is to implement, test, and validate tools that can be used to downscale seasonal climate forecasts into series of daily weather needed as input for process models (agronomic, hydrological) used to explore outcomes of diverse climate scenarios (Wilby et al. 1998; Corte-Real et al. 1999; Palutikof et al. 2002). The work involves two main objectives:

- The first objective is to implement and validate in the Southeastern United States a semi-parametric weather generator developed by Apipattanavis et al. (2007). This “hybrid” model addresses some of the weaknesses of previous weather generators by combining both parametric and nonparametric approaches. The hybrid generator will be used to generate daily weather series at several locations in the study region. Then, multiple diagnostic tests will be used to assess the performance of the generator by comparing properties of the historical and synthetic data.

- The second objective is to create sets of synthetic weather series that are consistent with (or conditioned on) particular climate scenarios (e.g., those predicted in seasonal climate outlooks). The simulated weather series, in turn, can provide input to process models, and through repeated simulations it would be possible to obtain frequency distributions of sectoral outcomes for each climate scenario. The conditioned synthetic series will be obtained by resampling the synthetic daily weather generated by
the semi-parametric generator implemented in the first part of the work. An approach will be developed to resample based on two different available formats of seasonal climate prediction. In the first format, seasonal climate forecasts state the likelihood of regional precipitation or mean temperature falling within certain categories (e.g., “above normal”, “below normal”). In the second format, the expected climate scenario is specified by the full probability distribution of a given variable (e.g., seasonal total precipitation).

This work focuses on three states in the Southeastern United States (SEUS): Alabama, Florida, and Georgia. This region has several important climate-sensitive sectors including agriculture, tourism, and water resources (Schmidt et al. 2000). Currently, crop simulation models are used in this region to predict the yield of some crops (e.g. Jones et al. 2003) (see http://www.agclimate.org). If climate information were incorporated into the weather sequence input for these models, agricultural stakeholders in the SEUS could assess the impact of expected variations in climate upon their interests (Baigorria et al. 2008). Research about crop simulation models and the incorporation of climate information into the weather input is currently being done by organizations such as the Southeastern Climate Consortium (http://secc.coaps.fsu.edu), an organization whose mission is to “provide scientifically sound information and decision support tools for agriculture, forestry, and water resource management in the Southeastern United States”.

1.4 The Study Area

The area of study in this project encompasses the states of Alabama, Florida, and Georgia in the SEUS. This area includes four different types of landforms (Burkett et al. 2001). The lower third of the state of Florida is in a sub-tropical flat with a significant
amount of marshland (or land that was once marshland). The rest of Florida and the lower portions of Alabama and Georgia are defined as being in a coastal plain. The upper portion of the states of Alabama and Georgia are considered part of the Piedmont plateau. Finally, a small portion of Northern Georgia is in the foothills of the Appalachian Mountain Range, at a relatively higher elevation than the rest of the study area (Burkett et al. 2001).

Most of the area of study has a temperate climate, with the lower portion of South Florida experiencing a tropical climate. Maximum and minimum temperature values increase from North to South in all seasons. Times of seasonal precipitation maxima and minima vary within the SEUS. The highest seasonal precipitation totals occur in southern Florida during the summer; while the lowest totals occur during the late fall and winter in southeastern Georgia and portions of the Florida Peninsula.

In this region, there are two primary types of atmospheric processes which drive precipitation. In late spring and summer, maritime tropical air steered by the Atlantic subtropical high pressure system leads to a high amount of convective activity in the SEUS. This convective activity is the main cause of precipitation in these months, although hurricane influences also may contribute significant amounts of precipitation in the late summer (Soule 1998; Baigorria et al. 2007). In other months, mid-latitude wave cyclones associated with fronts resulting from the interaction between warm, moist air from the Gulf of Mexico and cold, polar air masses are the primary cause of precipitation in the SEUS; the impact of these cyclones is smaller in southern Florida, where some fronts do not reach (Soule 1998; Baigorria et al. 2007). Precipitation due to frontal activity is strongest from November to March (Henry et al. 1994).
The SEUS is subject to interannual climate variability due to ENSO, whose signal is strongest in the fall and winter. Across the study area, La Niña falls and winters tend to be warmer and dryer than climatology, while El Niño falls and winters tend to be cooler and wetter than climatology (Ropelewski and Halpert 1987; Kiladis and Diaz 1989; Sittel 1994). La Niña summers tend to have higher precipitation totals in parts of the SEUS; however, this could be due to the increased number of tropical storms during La Niña phases (Bove et al. 1998). Peters et al. (2003) found that ENSO-related climate variability has a signal on vegetation conditions. They found that the neutral phase provides better conditions for vegetation, both cropland and forest, while the El Niño phase provides the least favorable conditions for vegetation. Harrison and Meindl (2001) proved a statistical relationship between ENSO and wildfire risk in Florida. They showed that wildfire activity increased in years following La Niña events, compared to years following El Niño events.

In Florida, El Niño falls and winters tend to be associated with higher river streamflow, while La Niña winters and springs are associated with lower streamflow levels (Schmidt et al. 2000). Research for the Tampa Bay region (FL) has shown that coastal water quality (fecal coliform levels) correlates with ENSO phase (Lipp et al. 2001). El Niño winters tended to have higher fecal coliform levels in this watershed, while La Niña winters tended to have lower fecal coliform levels. This is possibly due to the documented variations in river discharge and precipitation among ENSO phases (Lipp et al. 2001).

Agricultural production in the SEUS has been shown to be sensitive to climate variability. The SEUS is a major source of many fruits and vegetables for the United
States (Mearns 2003). Another important crop is cotton (Baigorria et al. 2008); this region provides 40% of the total value of cotton for the United States (Mearns 2003). Examples of other crops produced include wheat, peanuts, tobacco, corn, soybean, and sorghum (Burkett et al. 2001). For the SEUS, Hansen et al. (1998) found that the yields of maize and tobacco and the total value of maize, tobacco, soybean, and peanut were influenced by ENSO phase. In Florida winters, Hansen et al. (1999) found that yields of tomato, bell pepper, sweet corn, and snap bean were lower in El Niño years than during neutral or La Niña years. This could be due to higher precipitation amounts, lower daily maximum temperature, and decreased solar radiation during El Niño events (Hansen et al. 1999). Changes in cotton yield in the SEUS have been shown to be associated with climate variability caused by atmospheric circulation patterns in summer months (Baigorria et al. 2008). Decreased cotton yields may result from increased humidity, temperature, and rainfall (Baigorria et al. 2008).
Chapter 2
Data and Methods

This section describes in detail the data and approaches used in this work. First, the data utilized in this work are discussed. Next, a review of weather generator methods is presented, and the semi-parametric weather generator developed by Apipattanavis et al. (2007) to be used in this work is described in detail. This generator attempts to address some of the weaknesses of previous approaches by combining a parametric Markov model and a nonparametric k-NN procedure. Modifications made to the model to improve its performance are explained. Next, the approach developed for creating sets of synthetic series conditioned on seasonal climate forecasts is reviewed. This approach involved resampling the output of the semi-parametric weather generator, and was designed to work with forecasts in two different formats. One of the forecast formats consists of terciles and associated probabilities, and the resampling methodology requires knowledge of the values separating these terciles. The four methods that were utilized to calculate these values are introduced. Figure 2.1 provides a schematic of the methodology followed in this thesis.

2.1 Climate Data

The daily historical weather data used in this study were collected at cooperative weather stations and provided by the U.S. National Climate Data Center (NCDC, http://www.ncdc.noaa.gov). The data include 56 stations in Alabama, 90 stations in Florida, and 62 stations in Georgia. At each station, volunteer observers recorded daily
measurements of minimum temperature (Tmin), maximum temperature (Tmax), and total precipitation. Record lengths ranged from a minimum of 37 years (1 October 1969 to 30 September 2006) to a maximum of 67 years (1 October 1940 to 30 September 2006). The NCDC had performed quality control of these data sets.
Data for stations with more than 50% of missing daily values for any variable were removed before any work was done; out of the original 208 stations, 197 passed this criterion.

For the weather generator analysis, a subset of 11 stations was selected for detailed diagnostics: three in Alabama, four in Georgia, and four in Florida. Location of these stations is shown in Figure 2.2. The stations considered were chosen using various criteria, including coverage of different climatic regimes and the geographic range of the SEUS. Furthermore, stations selected typically are close to important areas of agricultural production. Agricultural products grown near stations in Alabama include peanut, cotton, maize, and timber. Products grown near stations in Florida include peanut, cotton, soybean, blueberry, vegetable, timber, and citrus crops. Products grown near stations in Georgia include peaches, peanuts, cotton, maize, pecans, onions, timber, and apples.

Table 2.1 lists the stations analyzed and provides information about their data. Record lengths for these stations ranged from 57 years (1 January 1950 to 31 December 2006) to 60 years (1 January 1947 to 31 December 2006). For each station, the percentage of daily records with any missing value ranged from 0.03% to 3.63%. Percentage of missing daily Tmin and Tmax values ranged from 0.01% to 2.05%. For daily precipitation values, percentage of data missing ranged from 0.01% to 2.91%. Only one station had more than 1.00% of missing precipitation values, and only three stations had more than 1.00% missing of any variable.

Some of the data sets used potentially have inherent quality issues due to the use of volunteers to collect data. Stations where volunteers were used include: Belle Mina (AL), Blairsville (GA), Brooklet (GA), Chipley (FL), Clermont (FL), Moore Haven (FL),
and Tifton (GA). All other locations are first-order stations at airports (operated by the National Weather Service): Atlanta (GA), Miami (FL), Mobile (FL), and Montgomery (AL).

For each daily record, values of the recorded weather variables correspond to the 24-hour period that precedes the time of observation (TOB). At airports, the TOB is midnight (local time), thus the observation period is the same as a calendar day. However, at volunteer stations, the TOB is often not midnight, thus the observed period

Figure 2.2: Map of stations used for generating synthetic series.
<table>
<thead>
<tr>
<th>Station</th>
<th>State</th>
<th>Lat (°)</th>
<th>Lon (°)</th>
<th>Elevation (m)</th>
<th>Record Length (Years)</th>
<th>% of Incomplete Daily Data Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atlanta</td>
<td>GA</td>
<td>33.63</td>
<td>-84.45</td>
<td>307.8</td>
<td>60</td>
<td>0.29</td>
</tr>
<tr>
<td>Belle Mina</td>
<td>AL</td>
<td>34.68</td>
<td>-86.88</td>
<td>182.9</td>
<td>57</td>
<td>1.27</td>
</tr>
<tr>
<td>Blairsville</td>
<td>GA</td>
<td>34.85</td>
<td>-83.95</td>
<td>594.1</td>
<td>60</td>
<td>0.64</td>
</tr>
<tr>
<td>Brooklet</td>
<td>GA</td>
<td>32.38</td>
<td>-81.67</td>
<td>54.9</td>
<td>60</td>
<td>0.56</td>
</tr>
<tr>
<td>Chipley</td>
<td>FL</td>
<td>30.78</td>
<td>-85.48</td>
<td>39.6</td>
<td>60</td>
<td>3.63</td>
</tr>
<tr>
<td>Clermont</td>
<td>FL</td>
<td>28.45</td>
<td>-81.75</td>
<td>33.5</td>
<td>58</td>
<td>1.88</td>
</tr>
<tr>
<td>Miami</td>
<td>FL</td>
<td>26.83</td>
<td>-81.08</td>
<td>8.8</td>
<td>59</td>
<td>0.03</td>
</tr>
<tr>
<td>Mobile</td>
<td>AL</td>
<td>32.30</td>
<td>-86.40</td>
<td>65.5</td>
<td>59</td>
<td>0.46</td>
</tr>
<tr>
<td>Montgomery</td>
<td>AL</td>
<td>30.68</td>
<td>-88.25</td>
<td>61.6</td>
<td>59</td>
<td>0.89</td>
</tr>
<tr>
<td>Moore Haven</td>
<td>FL</td>
<td>25.78</td>
<td>-80.32</td>
<td>10.7</td>
<td>60</td>
<td>0.73</td>
</tr>
<tr>
<td>Tifton</td>
<td>GA</td>
<td>31.45</td>
<td>-83.48</td>
<td>115.8</td>
<td>60</td>
<td>0.89</td>
</tr>
</tbody>
</table>

Table 2.1 Cooperative station data records.

(24 hours preceding the TOB) may be different from the calendar day. Volunteers prefer to take measurements in the early morning (e.g., 7:00 AM) or late afternoon (e.g., 5:00 PM) (Janis 2002), and this could lead to differences between the values of variables on the actual calendar day and the observation day. For example, if measurements are taken at 7:00 am, the minimum temperature in that 24 hour period could be from the previous morning; while for measurements taken at 5:00 pm, the highest Tmax may have occurred on the previous calendar day (Janis 2002).

There are other quality issues that arise through the use of volunteers. The TOB has changed throughout the record for some volunteer stations, with some locations having as many as five different TOBs. In addition, research has shown that volunteers
tend to favor values ending in “0”, “2”, “5”, and “8”. This is due to both human psychology and changing rules issued by the federal government (Nese and Dutton 1994).

2.2 Weather Generators

Weather generators are computer models that produce synthetic series of weather with the same statistical properties as historical data. Weather generators can be classified into two major types: those using parametric or nonparametric methods (Sharif and Burn 2006). Some of the simpler weather generators are used to generate only precipitation (stochastic precipitation models). However, many models generate additional variables such as minimum and maximum temperature, solar radiation, etc. This model focuses on generating precipitation, maximum temperature, and minimum temperature (i.e. the variables reported in the cooperative station data set).

2.2.1 Parametric Weather Generators

Parametric generation of synthetic series usually consists of a sequence of three major steps: (a) determining precipitation occurrence (i.e., defining the sequence of “dry” and rainy or “wet” days), (b) calculating precipitation intensity (the precipitation amounts on wet days), and (c) calculating the values of all other weather variables. For precipitation occurrence, a “wet day” is typically defined as a day having total precipitation greater than or equal to a designated threshold (e.g., one hundredth of an inch, HI), whereas a day having precipitation less than this threshold is considered a “dry” day.
A simple model of the precipitation occurrence process is a two-state, first-order Markov chain where the two states correspond to dry and wet days. In a first-order Markov model, the probability of the precipitation state on a given day depends only on the precipitation state on the previous day. The simple structure of a first-order Markov model enables the analytical determination of many of its properties, facilitating its subsequent application (Katz and Parlange 1998). This model has been used in many situations with remarkable success (Gabriel and Neumann 1962; Richardson 1981; many references in Hutchinson 1987; Wilks 1989; Katz and Parlange 1993). One limitation of the two-state, first-order Markov approach is the inaccurate simulation of long runs of dry or wet sequences (also known as clustering) at some locations (Racsko et al. 1991; Lettenmaier 1995; Semenov and Porter 1995).

When first-order models have been found inappropriate, an alternative has been to increase the order of the Markov chain (Jones and Thornton 1993; Katz and Parlange 1998). In higher-order models, probabilities are conditioned on the previous \( k (k > 1) \) days (where \( k \) is the order of the chain). Higher-order Markov chains in some cases have been found to simulate properties of precipitation (number of wet days) better than first-order chains (Katz and Parlange 1998). Katz and Parlange (1998) found that at some locations a higher-order Markov model reduced “overdispersion” (i.e., the tendency of weather generators to underestimate the observed variance of interannual monthly - or seasonal - total precipitation).

A disadvantage of higher-order Markov chains is that a larger number of parameters are required; a \( k \)-th-order, two-state Markov process is defined by \( 2^k \) parameters, and this number increases even more when generation is conditioned on
climate indices. In addition, parameter estimates for high-order Markov chains can be unreliable, especially when rainfall records are short (Hutchinson 1987). An alternative to these models is a hybrid-order Markov chain, which attempts to improve the simulation of dry spells only (Wilks 1999). In these models, the order of the Markov chain is increased only for simulation of dry spells.

Finally, as an alternative to Markov chains, spell-length models can be used to generate precipitation occurrence (Roldan and Woolhis 1982; Wilks 1999). These models fit probability distributions to the different spell length frequencies in the observed data. These models have been found to work well when other models have not modeled adequately the frequency of long dry spells. Types of distributions frequently used include the negative binomial, mixed geometric, and truncated negative binomial distributions (Wilks and Wilby 1999).

The next step is the generation of precipitation intensity. Here, different statistical distributions are fit to observed precipitation totals on wet days. Examples include the lognormal, cubic root normal, exponential, mixed exponential, kappa, gamma, and Weibull distributions (Bruhn et al. 1980; Richardson 1981; Stern and Coe 1984; Hutchinson 1987; Woolhis 1992). The most commonly used distribution for wet day amounts is the two-parameter gamma distribution (Wilks 1992; Husak et al. 2007). In some cases, the mixed exponential distribution has been shown to have the best fit to precipitation data, capturing extremes better than the gamma distribution (Wilks 1999).

In parametric generators, climate variables other than precipitation often are simulated using autoregressive models dependent on precipitation state. For example, the widely used WGEN model developed by Richardson (1981) relies on a first-order
autoregressive model conditioning weather variables on precipitation occurrence. In this model, all variables except daily precipitation are generated simultaneously, conditioned on the precipitation state of the current day.

Parametric stochastic weather generators are easy to implement, have a rich background, and have been successfully used in many situations. Some commonly used generators such as WGEN (Richardson 1981) are frequently included as modules of crop simulation packages (Pickering et al. 1994). Nevertheless, there are several drawbacks to parametric approaches to weather generation: (1) autoregressive models assume normal distributions of the variables and consequently, non-normal features in the data such as bimodality cannot be captured (Harmel et al. 2002); (2) only linear relationships between variables can be reproduced; (3) a large number of parameters have to be fitted for each season, which increases if simulations are to be conditioned on large scale climate indices; (4) by simulating other weather variables conditioned on precipitation, only that part of the dependency that is related to precipitation is captured; (5) lag-0 and lag-1 correlations of the variables are often not accurately reproduced; and (6) models used at one site may not be suitable for other locations (Rajagopalan and Lall 1999; Yates et al. 2003).

2.2.2 Non-Parametric Weather Generators

An alternative to parametric weather generators is the use of nonparametric models. Nonparametric models are data-driven and, unlike parametric methods, do not require assumptions about the distributions of the variables simulated. In addition, these models often are parsimonious, can capture non-linearities, and can easily be modified to do simulations for particular climate scenarios (Rajagopalan and Lall 1999; Yates et al.
Weaknesses of nonparametric methods include high computational requirements and limits to extreme value generation because of their reliance upon observed data (Wilks and Wilby 1999). That is, values that do not exist in the historical record cannot be generated.

There are several approaches to non-parametric weather generation. Use of empirical distributions of wet and dry spells and precipitation amounts is the simplest nonparametric approach, (e.g., the LARS-WG by Semenov and Porter 1995). Other approaches include the use of neural networks for the generation of temperature conditioned on global climate model output (Trigo and Palutikof 1999), and simulated annealing for generation of precipitation series (Bárdossy 1998). Rajagopalan et al. (1997) used a kernel-based method fitting a multivariate probability density function (PDF) to weather variables and, subsequently, simulating from the multivariate PDF. This approach has been applied successfully for monthly streamflow generation (Sharma et al. 1997) and daily weather generation (Rajagopalan et al. 1997), but it can be cumbersome for higher dimension problems (i.e., generating several climate variables simultaneously).

This work uses a non-parametric k-nearest neighbor algorithm (k-NN) (Rajagopalan and Lall 1999; Apipattanavis et al. 2007; Bannayan and Hoogenboom 2008). The k-NN method by Lall and Sharma (1996) involved the utilization of a weighted bootstrapping procedure to generate sequences of streamflow for a single location. Rajagopalan and Lall (1999) extended the k-NN method to multivariate data, simulating six weather variables. The k-NN bootstrap weather generator has subsequently been extended to multisite generation with good success (Buishand and Brandsma 2001;
Yates et al. 2003; Sharif and Burn 2006). Gangopadhyay et al. (2005) used the same model for downscaling ensemble weather forecasts at multiple sites. Furthermore, the k-NN approach was modified for resampling conditioned on atmospheric indices (Beersma and Buishand 2003) and hydrologic time series (Mehrotra and Sharma 2006).

Some of the advantages of the k-NN method include: (1) the dependence between all variables of interest can be well simulated, (2) lag-one correlations are often better captured than by other methods, and (3) they can be easily extended to multisite simulation (Yates et al. 2003). Advantage (1) results from the fact that all weather variables are selected together for a given day. Advantage (2) is assumed because all weather variables on any day \( t + 1 \) in the synthetic series are actual values in the historical data which follow historical values similar to the synthetic weather values of day \( t \). Advantage (3) is assumed since, in multisite simulation, the corresponding day’s weather is selected at all stations on any given day (Sharif and Burn 2006).

An issue with the k-NN generator is that it tends to under simulate the lengths of wet and dry spells (Aipattanavis et al. 2007). This behavior has been observed with other resampling approaches. For example, Young (1994) used diagnostics of his model that compared total number of synthetic and historical spells of different lengths (i.e. number of two-day dry spells, number of three-day dry spells, etc.) by month. Large differences between statistics of synthetic and historical data were found in certain situations with short monthly dry spell lengths. This underestimation of simulated spell lengths results from the intermittent nature of precipitation, and the fact that nearest neighbor selection often is based on all weather variables, including those that are not intermittent (e.g., maximum and minimum temperatures, solar radiation). Poor simulation
of spells can have significant impact in applications such as crop modeling, where sequences of wet and dry days are critical for modeling plant growth and, consequently, crop yields. To alleviate the spell problem, a modification to the k-NN weather generator was proposed by Apipattanavis et al. (2007).

2.3 The Semi-Parametric Weather Generator

The semi-parametric weather generator used in this work was developed by Apipattanavis et al. (2007) in an attempt to combine the advantages of parametric and nonparametric methods. The model includes a parametric Markov chain to simulate precipitation occurrence, and a nonparametric component based on the k-NN algorithm proposed by Rajagopalan and Lall (1999) to simulate values of all weather variables.

2.3.1 Markov Chain Model

To reproduce wet and dry spell statistics more accurately, Apipattanavis et al. (2007) modified the original k-NN resampling approach of Rajagopalan and Lall (1999) by introducing an additional step to simulate precipitation occurrence using a first-order, two-state Markov chain. The added Markov chain enhanced considerably the simulation of spell distributions. Nevertheless, diagnostics showed that annual total precipitation amounts still were underestimated. This deficit was related to a lower than observed frequency of medium-to-high daily precipitation amounts, as the overall number of rainy days was well described by the first-order, two-state Markov chain.

To address this issue, Apipattanavis et al. (2007) proposed a first-order, three-state Markov model. In this approach, precipitation is categorized into three states: (i) dry (daily precipitation < 0.3 millimeters or in this work, 1.18 hundredths of an inch),
(ii) very wet (precipitation greater than a specified percentile - in this work, the 80\textsuperscript{th} - of daily amounts for the month being simulated), and (iii) wet otherwise. Using three states, Apipattanavis et al. (2007) found that large daily precipitation amounts were simulated better, and annual precipitation totals (which were undersimulated by a two-state Markov chain) were reproduced better. Gregory et al. (1993) found that a first-order, many-state Markov chain can capture a high fraction of the seasonal variability, because the use of many states improves the model’s representation of spells of heavy precipitation which appears to have a considerable influence on the seasonal variance.

In the first-order Markov model, the assignment of precipitation occurrence is dependent only on the state of the previous day and is determined based on probabilities of transitions between any two states. The Markov model can be described by the equation:

\[ P_{ij} = \Pr \{ J_t = j \mid J_{t-1} = i \} , \quad \text{with} \quad i, j = 0, 1, 2 , \]

where \( P_{ij} \) is the probability of transition between states \( i \) and \( j \), which can take values of 0 (dry), 1 (wet) or 2 (very wet). As the Markov chain is first-order, the probability of precipitation state on day \( t, J_t \), is dependent only on the state of the previous day, \( J_{t-1} \).

The sequence of precipitation occurrence is determined by random selection of possible transitions between states based on their historical probability of occurrence. The estimation of transition probabilities is limited to days within a predefined temporal “window” (encompasses 15 days in this work) centered on the day of interest. For example, assume that the precipitation state “wet” is assigned to January 1\textsuperscript{st} of a series being simulated. To select the state of the following simulated day, the probability of all
transitions starting from a wet day (i.e., wet to dry, wet to wet, or wet to very wet) must be estimated. To estimate the transition probabilities, the number of occurrences of each transition type is counted for all appropriate two-day pairs in the historical record within a time window centered on the day of interest (January 1st). If a window of 15 days (seven days before and seven days after the day of interest) is used, then all two-day pairs are considered for which (a) the dates of the first days are between December 25th (seven days before January 1st) and January 8th (seven days after January 1st) and (b) the first day of the pair is wet. As an example, Table 2.2 provides a matrix of transition probabilities computed by the model for a three-state chain on January 1st at Chipley (FL).

<table>
<thead>
<tr>
<th>Transition Probabilities</th>
<th>Day One</th>
<th>Day Two</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dry Day</td>
<td>0.75</td>
<td>0.20</td>
</tr>
<tr>
<td>Wet Day</td>
<td>0.54</td>
<td>0.37</td>
</tr>
<tr>
<td>Very Wet Day</td>
<td>0.50</td>
<td>0.30</td>
</tr>
</tbody>
</table>

Table 2.2: Matrix of transition probabilities for a three-state Markov model.

The choice of the temporal window size for estimation of statistics includes tradeoffs between competing factors. A larger window increases the number of days available for calculation of statistics. It also provides more available candidate neighbors for the model, which can be important for simulating the less frequent very wet days. In some instances, the number of available very wet days may be very small, especially when this category is defined by relatively high daily precipitation thresholds (i.e. the 80th percentile of monthly distributions of daily precipitation) and historical records are short. However, as the window size increases, seasonal effects become a concern (Yates et al. 2003). For example, the characteristics of precipitation occurrence and intensity on March 1st in the historical data could be significantly different from the characteristics of
precipitation occurrence and intensity on March 20\textsuperscript{th}. In this work, a window of fifteen days (window encompassing seven days before and seven days after the day of interest) was adopted as a compromise between these competing factors. The same temporal “window” concept is used in the k-NN algorithm.

2.3.2 Nearest Neighbor Algorithm

Once the sequence of precipitation occurrence has been determined, the next step is to generate values for precipitation, maximum temperature, and minimum temperature. Before this can be done, seasonal effects are first removed from the historical data. To do this, the distributions of days on each calendar day in the historical data are standardized to an approximate mean of zero and standard deviation of one. This is done separately for dry days, wet days, and very wet days for all variables. To do this, means and standard deviations are calculated within the window of fifteen days spanning the calendar day in the historical data for each precipitation state. For each distribution of days (separated by precipitation state) on a given calendar day, the appropriate mean is subtracted from each value and then this value is divided by the appropriate standard deviation.

The daily values are selected by using a k-NN algorithm based on the model by Rajagopalan and Lall (1999). In this model, the k-NN algorithm uses a Euclidean distance measurement (see Equation 2.2 below) to determine the $k$ nearest neighbors to a given day in the synthetic series. The $k$ nearest neighbors are those days in the historical data most meteorologically similar in multivariate (precipitation, minimum temperature, maximum temperature) space to the current day (which is also referred to as the feature day) in the synthetic series. One of these $k$ neighbors is selected by resampling. The daily
values of the following day of the $k$ neighbor become the daily values for the following
day in that synthetic series.

A suitable value of $k$ can be the square root of the number of eligible neighbors.
Potential neighbors are days in the historical data within the specified window of fifteen
days encompassing the calendar day of the feature synthetic day. Eligible neighbors must
have the same precipitation state as that of the feature synthetic day and its following day
(or successor) in the historical data must have the same precipitation state as the
successor in the synthetic series. A step by step description is provided below and Figure
2.3 provides a schematic of the semi-parametric model of Apipatanavis et al. (2007).
The k-NN procedure used in this work is briefly described here:

1. Suppose the weather on synthetic day 1 (feature day), January 1\textsuperscript{st}, is $x_f$ (a vector of
temperature and precipitation values on the feature day) and the weather for target
day 2 needs to be simulated. The precipitation state of day 1 is wet and that of day
2 is dry (these states have been defined by the Markov process).

2. A 15-day window is centered on January 1\textsuperscript{st} (Dec 25\textsuperscript{th} – Jan 8\textsuperscript{th}) and all wet-dry
pairs of days within the window are selected from the historical data. The wet
days in each of these pairs are potential neighbors to the feature day, and the
values of one of the following dry days will become the simulated weather for
day 2.

3. Euclidean distances are calculated between the vector of synthetic values for day
1, $x_f$ and the weather variables for all potential neighbors in the historical data, $x_m$.
The Euclidean distance is of the form:
\[ r_{im} = \sqrt{\sum_{j=1}^{d} W_j (x_{ij} - x_{mj})^2}, \]  

(2.2)

where \( x_{ij} \) is the \( j \)th component (or data for variable \( j \)) and \( W_j \) is the weight assigned to variable \( j \). These weights are specified in the model and are the same (1.0) for all variables.

4. After the distances are computed, potential neighbors are ordered from the nearest (smallest Euclidean Distance) to the farthest (largest Euclidean Distance) and the closest \( k \) neighbors are selected where \( k \) is the square root of the number of available neighbors. Note that in some situations when the number of neighbors is low, for example with pairs involving very wet days, the value of \( k \) used can be different.

5. Each \( k \) neighbor is assigned a weight for resampling based on its Euclidean distance to the feature day. The closest neighbor (smallest distance) receives the largest weight and the furthest neighbor receives the smallest weight. Weights are defined by a discrete kernel \( K[j(i)] \):

\[ K[j(i)] = \frac{1/\sqrt{j}}{\sum_{j=1}^{k} 1/\sqrt{j}}. \]  

(2.3)

The element, \( j (i) \), is associated with the \( j \)th closest neighbor. The weights are normalized so that they sum to one, thus forming a probability metric.

6. One of the \( k \) neighbors is randomly selected using the weights calculated in the previous step. The daily weather of the dry day after the selected neighbor becomes the simulated weather for synthetic day 2.

Steps 1 through 6 are repeated to create synthetic series of the desired length.
Figure 2.3: Schematic diagram of the semi-parametric generator in Apipattanavis et al. (2007). (Reproduced from Apipattanavis et al. (2007) in Water Resources Research 2007, vol. 43). Scenarios are the terminology used for synthetic sequences.
2.3.3 Changes in the Calculation of Euclidean Distance

Initial testing of the hybrid generator for use in the SEUS showed that simulated annual precipitation totals were generally lower than the historical values. Different modifications to the weather generator algorithm were tested to improve simulation of precipitation. Two ways of calculating Euclidean distance for neighbor selection were tested: (a) using all weather variables in the historical record (i.e., Tmax, Tmin and precipitation) to calculate Euclidean distance between candidate neighbors and (b) using only Tmax and Tmin (i.e., excluding precipitation from the calculation of Euclidean distance). Diagnostics revealed that calculating Euclidean distance with method (b) (when only Tmax and Tmin were used) led to better simulation of annual precipitation totals.

While not using precipitation to select the closest neighbor to a given day seems to be using less information than is available, it was felt that this option was preferable due to superior simulation of precipitation statistics. As the Markov model uses three precipitation states, this implicitly recognizes the role of precipitation magnitude. That is, the three-state Markov model forced the generator to select high precipitation values, thus some information on precipitation intensity (by separation of precipitation into three states) was implicitly considered in the generation process. For these reasons, in all simulations described in this thesis, precipitation was not used in the Euclidean distance calculations.

2.3.4 Temperature Bias Corrections

Another problem detected during initial testing of the weather generator performance was a systematic bias in the grand means of Tmax and Tmin, i.e., the mean
of all daily Tmax and Tmin values in an entire synthetic weather sequence. This pattern was present at all locations.

Multiple attempts were made to understand and address this issue. Characteristics of the historical data, the behavior of the k-NN procedure, and parameterization of the weather generator were all issues that were explored. Some research revealed that interactions between the k-NN algorithm and the behavior of temperature in the historical data during transitions and during dry spells may contribute to the systematic bias, but no conclusive explanation was found for this problem.

To analyze this bias in more detail, first, monthly means of daily Tmax and Tmin were computed separately for wet and dry days for each synthetic series and for the historical record. Then, for each synthetic series, the magnitudes of deviations from historical temperatures were determined by subtracting the historical monthly means of daily Tmax and Tmin from the corresponding synthetic monthly means. Boxplots of these monthly deviations (indicating the dispersion of these values) for the 100 synthetic series generated at Atlanta (GA) are shown in Figures 2.4 and 2.5 respectively. Values in the boxplots lying above the horizontal line indicate a positive deviation (i.e., synthetic mean > historical mean), values lying below the horizontal line indicate a negative deviation (synthetic mean < historical mean). The medians of the boxplots (represented by the horizontal line within the boxes) are considered the monthly biases of the 100 synthetic series. Thus, if the median coincides with the horizontal line, for that month, there is zero bias in the synthetic data.
The signs of biases (medians of the boxplots) were opposite on dry days as compared to wet days. Figure 2.4 reveals that positive monthly biases were as large as +2 °F. Conversely, in Figure 2.5, on wet days monthly biases were as almost large as -2 °F. These biases had an annual cycle, with the largest biases (negative for wet days and positive for dry days) in the winter, and the smallest biases in the summer. Biases during summer months were much smaller, or even had an opposite sign as compared to other months.

Figure 2.4: Boxplots of the monthly average deviation (as an indicator of bias) of a) daily Tmax (°F) and b) daily Tmin (°F) for each synthetic series, calculated separately for each month of the year on dry days at Atlanta (GA). The boxplots indicate the dispersion of the deviations for the synthetic series. The horizontal line indicates zero deviation (or no bias).
To remove the temperature biases, additional code was inserted into the algorithm that allowed for the addition of empirical corrections to daily Tmax and Tmin values. These corrections are calculated and applied for each month and separately for wet and dry days. Corrections are equal to the medians of the deviations indicated by the horizontal lines within the boxplots in Figures 2.4 and 2.5. They are added empirically after the generation process to the original synthetic temperature values, and the new adjusted synthetic series are written to output files.
2.4 Conditioning Synthetic Series on Climate Forecasts

Process models could provide climate information at the scale of decisions (that preferred by stakeholders) if climate forecasts were incorporated into the synthetic weather input for these models. This could allow process models to provide more accurate simulations for decision makers, for example, by reducing the spread of likely outcomes. In regions (such as the SEUS) that are subject to interannual climate variability, the characteristics of precipitation can vary with different climate states. Non-climatological precipitation forecasts sometimes have been shown to have distributions of quarterly total precipitation that are shifted or narrower than the corresponding climatological distributions (for example, a shift downward in likely precipitation totals associated with a dry forecast) (Potgieter et al. 2003). The range of plausible process model outcomes using synthetic weather series that reflected these changes in the distribution of precipitation totals would enable decision makers to assess how their particular interests might be affected (Grondona et al. 2000).

To incorporate climate scenarios into process models, climate information and the input required by process models must be at the same spatial and temporal scale; however, these scales are often not the same. Climate scenarios are typically at a more coarse spatial and temporal scale. For example, climate forecasts are often provided at spatial scale equivalent to that of global circulation models and at a temporal scale of one to three months. In contrast, process models operate at the level of decisions and are concerned with daily, nonlinear interactions between weather, soil, and vegetation. Methods are needed to overcome this mismatch before climate forecasts can be incorporated into the synthetic weather input for process models.
2.4.1 Forecasting Climate

The modeling and forecasting of seasonal climate variability has improved significantly in recent years (Goddard et al. 2001). Changes in sea surface temperature (SST) in the tropical Pacific Ocean can influence the structure of the atmosphere in certain parts of the globe. Alterations to the structure of the atmosphere can lead to an increase in the likelihood of a particular regional climate condition occurring (Palmer and Anderson 1994). With the ability to predict changes in the atmosphere, the changes in the likelihood of a regional climate condition can be forecasted. Due to an increase in observational capabilities (satellites, moored and drifting buoys, etc.) and advances in understanding of ENSO, scientists have the ability to predict with some skill the changes in SST in the tropical Pacific Ocean and changes in regional climate. Currently, scientists are developing both dynamical and statistical forecasts of SST anomalies in the tropical Pacific Ocean, forecasts of ENSO events, and forecasts of regional seasonal climate.

There are statistical and dynamic methods of making seasonal SST forecasts. Forecasts of SST are usually predictions of deviations of SST averaged over certain regions of the tropical Pacific Ocean. Regions for which averaged SST anomalies are observed and predicted include the NiÑO3, NiÑO3.4, and NiÑO4 indices (Goddard et al. 2001). Statistical models often use observed SST, surface pressure, and wind stress data to make predictions of future SST (e.g. Van den Dool 1994; Knaff and Landsea 1997). Most statistical models use single or multivariate linear methods. Some models combine statistical methods with dynamical models into “hybrid” models for predicting SST. These models often use a statistical atmospheric model coupled to a dynamical ocean model.
Dynamic methods of forecasting SST rely on both intermediate and global circulation models. Intermediate models simulate only the region around the tropical Pacific Ocean and predict both SST and surface winds. Dynamical global circulation models (GCM) can model the atmosphere and/or oceans of the entire planet (Goddard et al. 2001). Sometimes atmospheric GCMs (AGCMs) and oceanic GCMs (OGCMs) are coupled to predict SST. In coupled GCMs, the model output includes the global state of the atmosphere, oceans, and land surface. Currently, the skills of forecasts of SST by dynamical and statistical methods are relatively similar based on studies of predictions up to and including the 1998 ENSO event (Barnston et al. 1999). However, due to the potential for increased complexity, it is expected that dynamic forecasts will eventually outperform statistical forecasts of SST.

Scientists also are able to make dynamical and statistical forecasts of seasonal regional climate. Traditionally, predicted SSTs are used as the only predictor for statistical models of climate prediction; sometimes ENSO trends and atmospheric variables are used as additional predictors. Auto-regressive, (e.g. Elfandy et al. 1994), multiple linear regressive, and probabilistic (e.g. Mason and Mimmack 2001) methods are the most common statistical methods used.

There are two types of numerical models used for making dynamical climate forecasts: coupled models and AGCMs. AGCMs are used in the “two-tiered” approach to climate forecasting, where creating forecasts of climate and forecasts of SST are done separately. In this approach, one model predicts SST and then the AGCM takes those predictions and forecasts regional climate only. Coupled GCMs can be used in a “one-tiered” approach, in other words, both SST and regional climate are predicted by the
same model. Currently, coupled GCMs are more often used for SST prediction only, while AGCMs are more commonly used for climate prediction (Hunt 1997; Mason et al. 1999).

### 2.4.2 Climate Forecast Formats

The formats of seasonal climate forecasts range from the prediction of an El Niño or La Niña event (that in turn influence regional climate conditions) to probabilistic statements about the likelihood of regional precipitation or temperature falling within certain categories (e.g., “above normal” conditions). Seasonal climate forecasts are being operationally and experimentally disseminated by various agencies around the world, including the United States Climate Prediction Center (CPC, see http://www.cpc.noaa.gov) and the International Research Institute for Climate and Society (IRI, see http://iri.columbia.edu).

The most commonly used format in CPC and IRI forecasts displays the probabilities of total precipitation or mean surface air temperature falling in the lower (“below normal”), middle (“near normal”), or upper (“above normal”) thirds (or terciles) of the historical distribution of that variable for a particular region (Goddard et al. 2003). A forecast of “climatology” contains no information beyond the expected equal chance (1/3, or 33.3%) of the outcome being in each tercile of the distribution. In contrast, for a non-climatological forecast (e.g., when climate anomalies associated with an extreme ENSO phase are expected), the probabilities assigned to the terciles reflect the increased or decreased likelihood of that outcome. For example, a hypothetical forecast indicating 17%, 33%, and 50% probabilities of rainfall in the lower, middle, and upper terciles,
respectively, implies greater expectation (50% vs. 33%) of above-normal precipitation. In most forecasts, the probability assigned to the near-normal or middle tercile remains near 33%.

IRI global forecasts for temperature and precipitation are provided in tercile format for three-month periods and are updated monthly. The geographic scale of the models used for these forecasts is approximately 2.8° longitude and latitude. On the IRI web site, eight regional maps of these probabilities are provided which cover the entire globe; a global map also is provided. An example of a regional IRI forecast is provided in Figure 2.6. This is a precipitation forecast for March, April, and May (MAM) of 2008 for North and Central America. The yellow color over the SEUS indicates a “dry” forecast for the study area since the highest probability is assigned to the lower tercile (yellow indicates about 40% probability for the “Below-Normal” precipitation).

The CPC provides forecast maps for temperature and precipitation for the United States and surrounding areas only, and these forecasts are provided for both one-month and three-month periods. The maps do not provide numerical tercile probabilities, but note whether normal, above normal, or below normal conditions are expected. In addition to these maps, the CPC provides “probability of exceedance” forecasts which indicate the probability of exceeding a given value of a weather variable (e.g. seasonal precipitation total, monthly mean air temperature) for 102 climate divisions in the United States. These forecasts provide various useful numerical threshold values, including tercile probabilities and boundaries. An example of a probability of exceedance precipitation forecast for the South Florida climate division is provided in Figure 2.7. This forecast is issued for the same period (MAM 2008) as the IRI forecast in Figure 2.6. The region for
which this forecast (southern Florida) is provided is roughly similar in size to a single grid space in the IRI forecast, as the climate division shown covers approximately the lower half of the Florida Peninsula. Forecast probabilities for terciles are located in the figure above and to the right of the curve, and are identified as “Prob of above average” (28.5%), “Prob of near average” (33.1%), and “Prob of below average” (38.3%) for the
upper, middle, and lower terciles respectively. Since it is a dry forecast, the forecast curve is shifted to the left of the curve representing climatology (fitted to the historical data).

Noticeably, more information can be discerned for Southern Florida from this forecast as compared to the IRI tercile forecast in Figure 2.6. For example, multiple threshold values of quarterly total precipitation and associated probabilities can be discerned from the probability curve, while the tercile forecast provides a limit of three probabilities for the terciles and no numerical precipitation values.
2.4.3 Conditioning Synthetic Series with Weather Generators

Stochastic weather generators typically produce sets of synthetic series that together reflect the climatology of the weather data utilized as input. However, sometimes it would be preferable to incorporate climate information into these sets of synthetic series (i.e. for input into process models to assess impacts from climate variability). If seasonal climate forecasts were available, weather generators could be used to produce sets of synthetic daily series conditioned on the forecasts, thus generating sets of series reflecting predicted deviations from climatology.

Different approaches have been developed for conditioning synthetic series with the use of weather generators and they can be classified into three groups:

1. conditioning the parameters of weather generators,
2. resampling the input (i.e. historical data) to weather generators or
3. sampling the synthetic output based on the climate scenario of interest.

The most widely used method of conditioning synthetic series with generators involves adjusting the parameters of parametric weather generators to match particular statistics (such as monthly means). Examples include conditioning the parameters of weather generators on climate forecasts (Wilks 2002), ENSO phases (Grondona et al. 2000), indices of large scale atmosphere-ocean circulation (Katz and Parlange 1993; Katz et al. 2003), and climate change scenarios (Katz 1996; Mearns et al. 1997). Some weaknesses of using parametric generators for conditioning are the same as when doing parametric unconditional simulations; for example, these methods require assumptions about the distributions. Hansen and Ines (2005) noted that the relative contributions of changes in precipitation intensity and occurrence associated with fluctuations in climate variables (such as monthly precipitation) must be assumed. In addition, a larger number of parameters are needed (and must be estimated) when
conditioning. Finally, Katz (1996) found that adjusting the parameters for one variable may unexpectedly influence another variable due to their interdependence, and this may have unwanted influences. He also proposed a possible means of overcoming this obstacle.

A second method that has been utilized for producing series conditioned upon climate information more recently is resampling weather generator input based upon climate information. Apipattanavis et al. (2007) separated historical data files into wet, dry, and near normal yearly categories based on precipitation totals and resampled those categories based on probabilistic seasonal forecasts to create new input files. The generation of synthetic series subsequently was performed with a k-NN method using these modified input files. Another example of this approach involves ranking years (or weeks) in the input historical record based on prescribed climate change criteria (i.e. warmer or colder years). In this method, new input files were created resampling based on these rankings and the k-NN generation process was performed using these new weighted input files (Yates et al. 2003; Sharif and Burn 2006). Clark et al. (2004) used ranking within the input data for generating series conditioned on an ENSO index (an index of sea surface temperature anomalies in the central Pacific Ocean). In their work, they added an additional level of randomization to the resampling procedure where new modified input files were created for generating each synthetic day in the k-NN process.

Clark et al. (2004) noted that a weakness of this resampling method when used in climate change studies is that biased resampling favors sections of distributions. This can lead to changes in the characteristics of distributions, such as standard deviations and skewness, which may not be desirable when creating changing climate scenarios. Clark et
al. (2004) also noted that, in some situations, dependence on a small subset of data may lead to unwanted influence by extreme values (for example oversampling of large precipitation values) whose impact would be negated or reduced if the entire data set were used.

Examples of the third method (resampling weather generator output) include works by Hansen and Indeje (2004) and Hansen and Ines (2005). They modified the synthetic output from weather generators to create conditioned synthetic series. Here, the weather generator method (parametric or nonparametric) will not affect the conditioning performance as the only necessary ingredient is a large number of synthetic series. In these works, synthetic series were generated until target values were approximated; then a rescaling factor was used to match the synthetic statistics to targets exactly. Hansen and Ines (2005) compared this method with conditioning by parameterization of weather generators and noted that this method has some advantages: constraining the output was more efficient in terms of number of realizations required to achieve desired statistics. When constraining the output, no assumptions were required about the precipitation occurrence and intensity processes, thereby preserving the frequency and intensity relationships. Modification of parameters in parametric generators can lead to inaccurate intensity and occurrence relationships, which can significantly impact the modeling of crop yields. They found that in conjunction with crop simulation models, conditioning of parameters tended to under predict crop yields more than constraining output.

2.4.4 A New Approach for Conditioning Synthetic Series with the Semi-Parametric Weather Generator

A new approach was developed to create synthetic series conditioned on seasonal climate forecasts. The approach involved weighted resampling with replacement of
synthetic output from the semi-parametric generator based on seasonal precipitation forecasts.

This approach has some differences with previous conditioning approaches. Unlike approaches in groups (1) and (2), only the synthetic output of the generator is altered here, while the weather generator and historical input are unmodified. Thus, this approach avoids some of the potential problems that have been encountered when working with the input (Clark et al. 2004) or when modifying the parameters in parametric models (Hansen and Ines 2005). This method is more similar to type (3) in that only the synthetic output is used; however there are some differences between this work and previously discussed approaches of type (3). In this approach, the entire set of synthetic series created by the generator can be utilized, unlike in Hansen and Ines (2005) and Hansen and Indeje (2004), where only a subset of the output is used. Unlike this procedure, those works also involved altering the synthetic precipitation output. The synthetic daily precipitation values were multiplied by a scaling factor of different magnitudes for each month.

The resampling procedure was developed for utilizing two different forecast formats. The first format consists of the tercile probabilistic format (format I) (see Figure 2.6). With format I, conditioning is done by resampling based on the three tercile probabilities. The second format is in the form of a probability distribution function (format II) (see Figure 2.7), similar to the probability of exceedance forecasts provided by the CPC. Resampling based on format II is done by first, partitioning the probability distribution function (PDF) into a predefined number of equal-width bins. Then, the
relative density in each bin is determined, and these density estimates are converted to probabilities.

To test the resampling procedure, hypothetical seasonal forecasts were created using available historical IRI forecasts for the study area (http://portal.iri.columbia.edu/portal/server.pt). Five different forecast scenarios were constructed (shown in Table 2.3) for both formats: very dry, moderately dry, climatological, moderately wet, and very wet. The climate forecasts were constructed for non-overlapping quarters: January/February/March (JFM), April/May/June (AMJ), July/August/September (JAS) and October/November/December (OND). Of the available historical forecasts from the IRI, the most non-climatological forecasts were in JFM. The largest deviations from climatology in forecasts for AMJ and JAS were significantly less. For each three-month period, the most extreme forecast probabilities available from the IRI website for the study area were assigned to the very wet and very dry scenarios. Less extreme forecast probabilities were used for JFM moderately wet, JFM moderately dry, and OND moderately wet forecasts. For all other quarters and conditions (wet or dry), the same forecast probabilities were used for both moderately wet (moderately dry) and very wet (very dry) scenarios. This was done because only one set of non-climatological forecast probabilities (25%-35%-40% for wet forecasts or 40%-35%-25% for dry forecasts) was issued for these three month periods. The various scenarios are shown in Table 2.3. The lower tercile probabilities range from 50% (very dry for JFM) to 20% (very wet for JFM) while the upper tercile probabilities ranges from 15% (very dry for JFM) to 60% (very wet for JFM). The middle tercile probabilities stay near 33% in most forecasts.
2.4.4.1 Resampling Using Format I

Format I is the tercile forecast format, where probabilities are provided for three equal-width sections of the climatological distribution. To create conditioned synthetic series using this format, first the precipitation values of the daily synthetic series were summed over three month periods (the same as those of the forecasts) to create synthetic quarterly precipitation totals. The new distribution of synthetic quarterly totals was separated into terciles based on boundaries derived from the historical data (methods of calculation are discussed in the section 2.5). Resampling of the quarterly totals by tercile was performed based on the forecast tercile probabilities following the steps outlined below. This procedure was performed for all quarters and forecast scenarios.

<table>
<thead>
<tr>
<th></th>
<th>JFM</th>
<th>AMJ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lower</td>
<td>Middle</td>
</tr>
<tr>
<td>Very Dry</td>
<td>50%</td>
<td>35%</td>
</tr>
<tr>
<td>Moderately Dry</td>
<td>45%</td>
<td>35%</td>
</tr>
<tr>
<td>Climatological</td>
<td>33%</td>
<td>33%</td>
</tr>
<tr>
<td>Moderately Wet</td>
<td>20%</td>
<td>35%</td>
</tr>
<tr>
<td>Very Wet</td>
<td>20%</td>
<td>20%</td>
</tr>
<tr>
<td></td>
<td>Lower</td>
<td>Middle</td>
</tr>
<tr>
<td>Very Dry</td>
<td>40%</td>
<td>35%</td>
</tr>
<tr>
<td>Moderately Dry</td>
<td>40%</td>
<td>35%</td>
</tr>
<tr>
<td>Climatological</td>
<td>33%</td>
<td>33%</td>
</tr>
<tr>
<td>Moderately Wet</td>
<td>25%</td>
<td>35%</td>
</tr>
<tr>
<td>Very Wet</td>
<td>25%</td>
<td>35%</td>
</tr>
</tbody>
</table>

Table 2.3: Hypothetical forecasts for the different scenarios and quarters. In each forecast, the listed probabilities are assigned to the terciles.
Two different illustrations are provided for resampling with format I. In Figure 2.8, a schematic of the procedure is provided. Below, a step by step description is provided to describe the resampling process for one quarter (JFM).

The steps of this resampling procedure are:

1. Historical precipitation daily data in every JFM quarter was summed creating a distribution of historical quarterly precipitation totals (e.g. totals for JFM 1960, JFM 1961, etc.).

2. Boundaries which separate the terciles of the distribution of historical quarterly precipitation for JFM were estimated using bootstrapping approaches (see section 2.5).

3. Daily precipitation values for JFM of every synthetic year in the weather generator output were summed into corresponding quarterly totals.

4. Synthetic JFM quarterly precipitation totals were categorized into upper, middle, or lower terciles using the tercile boundaries from historical data estimated in step 2.

5. Resampling with replacement of the synthetic JFM totals in each tercile was performed based on the tercile probabilities from the hypothetical forecast of interest.

6. The daily data corresponding to each resampled quarterly total was extracted from the original synthetic series and placed randomly into one JFM quarter of a new sequence of synthetic daily data.

Steps 1 through 6 were repeated for each quarter and each forecast scenario.
Figure 2.9 provides histograms illustrating the changes in a data set resulting from this procedure. The first histogram corresponds to JFM quarterly total precipitation from the original historical data for Chipley (FL). The second histogram shows the synthetic JFM quarterly total precipitation from the conditioned set of synthetic series after the resampling procedure has been performed based on a hypothetical very dry forecast in
format I. The shift in the new distribution to lower values indicates that the procedure may change the properties of the synthetic series.

2.4.4.2 Resampling Using Format II

This procedure focuses on the entire distribution of a precipitation forecast instead of the three tercile probabilities. Here, weighted resampling was performed based on the probability density of the forecast distribution.

To use the same forecast scenarios for format II, forecast distributions were created from the historical data based on the tercile forecast probabilities. First, historical quarterly (JFM, AMJ, etc.) precipitation totals were computed from the daily data. Then,
the totals were resampled using the historical tercile boundaries and forecast tercile probabilities. The distribution of the resampled historical quarterly totals should have tercile probabilities equal to those of the corresponding forecast. For example, if the tercile probabilities associated with a non-climatological forecast were 50%-30%-20%, then, in the distribution of resampled historical quarterly totals (resampled based on this forecast), 50% of the precipitation totals should have values less than the lower tercile boundary (thus being in the lower tercile) while 30% of the precipitation totals should have values that would be categorized as in the middle tercile, and 20% of the resampled totals should be associated with the upper tercile. The density of the new set of historical resampled quarterly values was estimated using kernel density estimation (Bowman and Azzalini 1997). The result of this process was a PDF representing the forecast scenario.

The resampling procedure used for conditioning is similar to that used with format I; however, instead of separating the forecast into three terciles, the PDF forecast was partitioned into finer intervals of 20 hundredths of an inch (HI). After the PDF forecast was partitioned, the relative proportion of the forecast distribution was determined for each bin. Then, all of the relative proportions were converted into probabilities summing to 1.0. Next, for the same three month period, the distribution of synthetic quarterly totals was portioned using the same intervals. Resampling was performed based on the probability for each bin derived from the forecast. Resampling the synthetic output based on format II should allow for the capture of more detail of the forecast as compared to resampling based on format I, as a PDF can be subset at a much finer resolution than three sections (the terciles of format I).

When using forecasts in format II, resampling from the synthetic output may be
preferable than resampling from historical input files (methods of type 2). From a large number of synthetic series from the generator, the resulting dense distributions of variables can easily be divided into finer intervals associated with the bins of the PDF forecasts. Resampling from this data set may be preferable to resampling from a short historical data record (which would be used in methods of type 2) subset at multiple intervals. For example, a historical record of 60 years (as used in this work) would have 60 quarterly values. If the distribution were subset into 15 equal-sized bins, it is likely that several of the bins would have zero values.

Figure 2.10 provides a schematic of this procedure and a step-by-step description is provided below.

The steps of this resampling procedure are:

1. Historical precipitation daily data in every JFM quarter was summed creating a distribution of quarterly totals (e.g. totals for JFM 1960, JFM 1961,…).

2. A forecast distribution was created by separating this distribution of historical values by terciles and then resampling these historical quarterly totals based on the forecast tercile probabilities.

3. A PDF was estimated from the forecast distribution.

4. The forecast PDF was separated into 20 HI bins.

5. Density estimates were made for each bin.

6. These density estimates were converted into probabilities.

7. Daily series for JFM of every synthetic year from the weather generator output were summed into corresponding quarterly totals.
8. The distribution of synthetic JFM quarterly totals were separated into 20 HI bins equivalent to that of the forecast bins.

9. Forecasts probabilities were assigned to each quarterly total within each corresponding bin.

10. Resampling with replacement of the synthetic JFM totals was performed based on these assigned probabilities.

11. The daily data corresponding to each resampled quarterly total was extracted from the original synthetic series and placed randomly into one JFM quarter of a new sequence of synthetic daily data.

Steps 1 through 11 are repeated for each quarter and each forecast scenario.

2.5 Estimation of Tercile Boundaries

The resampling methodology described for format I requires knowledge of the boundaries between the terciles of climatological distributions of quarterly total precipitation and average temperature indicated by the 0.33 and 0.66 quantiles (hereafter Q33 and Q66). These values are specific to the location where synthetic series are generated, and at some locations, these values may not be readily available. Several methods are available in the literature for calculation of quantiles and the various methods may produce different estimates. Parrish (1990) studied the estimation of quantiles for Gaussian distributions of small sample size using nonparametric methods. Parrish noted that in some cases, the choice of statistical method for estimation of quantiles should be carefully considered; he found significantly different results for some quantile estimates when using different methods. Parrish noted that the type of parent distribution and sample size are factors that can contribute to the divergence among
estimates by different methods. The data sets that Parrish (1990) used were Gaussian; however, distributions of climate variables used in this work often are not Gaussian. These distributions have irregular characteristics such as bimodality, skewness, and presence of outliers. In addition, the sizes of these historical data sets are limited, and
climatological distributions of weather data can have random gaps due to missing or sparse data. Various methods may respond to these irregular characteristics differently, yielding diverging estimates of some quantiles.

In this work, various approaches to estimating Q33 and Q66 are explored: these are based on empirical estimates and on fitting both parametric and non-parametric distributions to the historical data. Large differences among estimates from the various methods and their likely causes were explored. In addition, historical data sets with varying simulated percentages of missing values were used to test the impact of incomplete series upon the variability of quantile estimates.

2.5.1 Data Preparation

At all 197 locations, Q33 and Q66 were estimated for distributions of seasonal total precipitation and average temperature (Tavg). Average temperature was calculated as the average of daily Tmax and daily Tmin. Daily precipitation and Tavg were aggregated into quarterly values: January/February/March (JFM), April/May/June (AMJ), July/August/September (JAS), and October/November/December (OND). For a Tavg (precipitation) quarterly value to be valid, the corresponding quarter had to have less than 15 (9) missing daily records. To estimate Q33 or Q66 for a given station, a minimum of 30 non-missing quarterly values were required for that variable.

2.5.2 Empirical Quantile Estimation (EMP)

The simplest approach to estimating quantiles was based on sorting the values of a given historical series (one station, one quarter, one climate variable) and deriving the values by linearly interpolating between order statistics of the target historical series, \( x \), assuming that the \( i \)th order statistic was the \( (i - 1)/(\text{length}(x) - 1) \) quantile. For
example, for a historical series of 6 sorted values (10, 12, 14, 16, 18, and 20), the 4th order statistic (16) is the (4-1) / (6-1) = 0.60 quantile.

To prevent empirical quantile estimates from being influenced by unusual values in the data set, bootstrap sampling (Efron and Tibshirani 1993) was used. Briefly, bootstrapped distributions of the same size as the original record are generated by resampling with replacement from the original data set (Dunn 2001). Because samples are drawn with replacement, some of the original values may appear more than once or not at all in each bootstrapped distribution.

For a given historical series, 1000 bootstrapped distributions were created. Next, for each of these 1000 bootstrapped distributions, the target quantiles were estimated using the empirical method discussed above. Finally, the average of these quantile estimates (one from each distribution) became the “bootstrapped quantile” and was used in subsequent analyses.

2.5.3 Estimation by Fitting a Gamma Distribution (GAMMA)

The second approach to estimating quantiles involved fitting a gamma distribution to each historical series. The versatility of this distribution makes it attractive for representing precipitation and often is used for this purpose (Wilks 2006). The gamma distribution is defined by two parameters: \( \alpha \) (the so-called shape parameter) and \( \beta \) (the scale parameter). Parameter values for each historical series were estimated using a robust approach (Marazzi and Ruffieux 1999), which minimizes unwanted influence by outliers. Once the distribution was fitted to the historical values, the inverse of the cumulative gamma density function was used to derive the target quantiles. As with the
previous method, 1000 bootstrapped distributions were generated and the target quantiles were estimated from each series.

2.5.4 Nonparametric Kernel Density Estimation (KERNEL)

The third approach involved fitting a non-parametric probability density function to each historical series. Because precipitation series are bounded at zero millimeters, a log-transformation was applied to precipitation data prior to density estimation to avoid undesirable effects near the lower boundary (Bowman and Azzalini 1997). In this method, no predetermined functional form (e.g., a gamma) was assumed. Instead, an empirical kernel density function:

\[ \hat{f}(x) = \frac{1}{nb} \sum_{j=1}^{n} K \left( \frac{x - x_j}{b} \right), \]  

(2.4)

was fit to each historical series of size \( n \) using a fixed Gaussian kernel \( K(\cdot) \), and bandwidth \( b \). The corresponding empirical cumulative density function was derived, and quantile values were calculated.

For kernel density estimation, the shape of the kernel is not critical to the estimates, but the choice of bandwidth is (Bowman and Azzalini 1997; Venables and Ripley 2002). The bandwidth selection represents a compromise between smoothing enough to remove insignificant bumps, but not smoothing too much to smear out real peaks or modes (Venables and Ripley 2002). Sheather (2004) reviews procedures to select appropriate bandwidth values. In this work, the plug-in method of Sheather and Jones (1991) was used. In the plug-in method, a bandwidth is selected to minimize estimates of the mean integrated squared error. The implementation used was the library
“sm” (Bowman and Azzalini 1997) available for S-Plus statistical software (see http://www.stats.gla.ac.uk/~adrian/sm/).

### 2.5.5 Density Estimation Using Adaptive Splines (SPLINE)

The final approach tested was adaptive estimation of a probability density function using B-splines, a type of basis function. As in the previous method, no functional form was assumed for a historical series. Different B-spline functions are applied to separate parts of the distribution to capture better its density shape. Following Kooperberg and Stone (1991, 1992), a log-density function was modeled as a cubic spline:

\[
\hat{f}(x, \theta) = \exp \left( \sum_{i=1}^{p} \theta_i B_i(x) - c(\theta) \right),
\]

where \( c(\theta) \) is a normalization factor, \( \hat{\theta}_i \) is the maximum-likelihood estimate, and \( B_i(x) \) denote the B-spline functions.

For each historical series, a density was fit using the adaptive splines method. Briefly, the algorithm begins by adding basis functions while searching for the greatest increase in the value of the log-likelihood; then deleting basis functions while searching for smallest decrease in the log-likelihood. Target quantile values were then directly estimated using the S-Plus BEST software library (see http://www.insightful.com/downloads/libraries). For precipitation, a lower bound of zero was defined to improve the density estimates.

The goal of the algorithm is to select a model that has the smallest value of Akaike’s Information Criterion (AIC). Choosing the number and location of knots is the main problem in adaptive spline density estimation; a small number of knots give an
estimate that is missing important details, while too many knots may produce a jagged fit. The number of knots, the location of the knots, and the types of B-splines can be specified or automatically selected during the computational process.

In some cases, this method produced curves that fit the data distribution poorly; the curves were too smooth and unrealistic. Some of these fits were examined and they all had three knots, the minimum allowed by the algorithm. It was hypothesized that the poor fits were due to a combination of small sample size (between 35 and 64 values) and the irregular characteristics of historical distributions. This was tested for a few cases by progressively increasing the number of records in the data set; doubling the sample size generally led to a more appropriate fit.

In an attempt to address the poor spline fits, limits were placed upon the adaptive algorithm using options available in the S-Plus BEST software. Kooperberg and Stone (1991) discussed the specification of a fixed number of knots, recommending different numbers depending upon the size of the data set. The lowest samples they studied had between 60 and 124 values, for which they recommended six knots. These recommendations were based on the assumption of distributions with unimodal shapes. Since our sample sizes ranged from 35 to 64 values, but many of the distributions were irregular, six knots were selected for estimating a second set of spline fits. The new fits were no longer overly smoothed. In the results, the original quantile estimates are discussed unless explicitly stated otherwise.
Chapter 3

Results

The performance of the tools used in this work is reviewed in this chapter. First, synthetic series of maximum temperature, minimum temperature and precipitation were generated from the semi-parametric weather generator at 11 different locations in the SEUS. The synthetic series were then examined to determine how well statistics of the historical data were replicated. Multiple diagnostics were used to assess reproduction of characteristics of daily weather and aggregate totals. The performance of the Markov chain and k-NN process was discussed. In addition, temporal patterns of climate variables were discussed within the context of the SEUS climate.

Next, the boundaries separating the terciles (0.33 and 0.66 quantiles) in climate forecasts needed for the resampling procedure were calculated at all stations using four different methods. The resulting estimates by different methods were compared. Instances of large differences among estimates were determined, and the causes of variability were explored. Maps of the quantile estimates for quarterly average temperature and total precipitation for the study area are provided.

Finally, using the resampling approach, conditional sets of synthetic series were created based on different precipitation forecast scenarios. The resulting conditioned series were analyzed to determine if they were consistent with the forecast scenarios. Distributions of quarterly total precipitation from the conditioned synthetic series were compared to corresponding historical distributions to check for expected deviations from climatology. In addition, diagnostics were used to assess if the properties of daily precipitation had changed.
3.1 Synthetic Output from the Weather Generator

At each location, 100 synthetic series – each the same length as the corresponding historical data (57-60 years) – were generated using the approach by Apipattanavis et al. (2007). Graphic diagnostics were then used to assess the model’s ability to generate daily sequences of precipitation, maximum temperature (Tmax), and minimum temperature (Tmin) statistically similar to the historical sequence. If the generator performed satisfactorily, the historical sequence could be considered a possible realization of the generator process. In this work, more extensive testing was done than traditionally performed in weather generator assessment (e.g. Rajagopalan and Lall 1999; Yates et al. 2003; Sharif and Burn 2006): 36 different graphic diagnostics were used. In the following sections, the results of a selected subset of diagnostics are shown and discussed.

3.1.1 Temperature Bias

In initial testing, the synthetic Tmax and Tmin values showed a consistent bias when compared to the historical data. On dry days, mean daily temperature values were consistently higher than the mean historical temperature values, whereas the opposite was true for wet days (see Figures 2.4 and Figures 2.5). Multiple attempts were made to understand and address this issue. Characteristics of the historical data such as skewness, elements of the k-NN procedure (including the standardization process), and the parameterization of the weather generator were all considered as potential causes of the bias and were analyzed in detail. Nevertheless, no conclusive explanation was found for this issue.

In the final model, empirical corrections were added to the daily synthetic Tmax and Tmin values to remove the bias. The corrections were estimated for each station
separately for wet and dry days on a monthly basis. For Tmax on dry days most corrections were negative, ranging from \(-2.53\) °F in February at Belle Mina (AL) to \(+0.78\) °F in July at Atlanta (GA). For Tmax on wet days most corrections were positive, ranging from \(-0.49\) °F in September at Montgomery (AL) to \(+1.31\) °F in December at Blairsville (GA). For Tmin on dry days most corrections were negative, they ranged from \(-2.69\) °F in February at Belle Mina (AL) to \(+0.33\) °F in July at Atlanta (GA). For Tmin on wet days most corrections were positive, ranging from \(-0.05\) °F in March at Mobile (AL) to \(+1.34\) °F in December at Blairsville (GA).

3.1.2 Assessment of Maximum Temperature and Minimum Temperature Generation

In the following section, the results of the analysis of the de-biased synthetic Tmax and Tmin values are reviewed. In each subsection, the diagnostic used is introduced, and the results are briefly discussed. In addition, the characteristics of temperature in the SEUS climate were characterized by describing the annual patterns of temperature statistics across different stations.

Boxplots were used in several graphic diagnostics discussed in this section (see Figure 3.1). The boxplots display the dispersion of the statistics (or values) of interest for the 100 synthetic sequences. The box represents the central 50% of the values and its width is referred to as the interquartile range. The horizontal line within the box indicates the median value. The vertical lines above and below the box (called whiskers) extend outward to 1.5 times the interquartile range of the data. The region above the interquartile range is called the upper quartile and the region below the interquartile range is the lower quartile. The whiskers and interquartile range together cover what is referred to as the
contiguous range of the boxplots. Any values outside the contiguous range are considered outliers and are indicated by separate horizontal lines.

3.1.2.1 Monthly Means of Tmax and Tmin

To assess the presence of temperature bias after the application of empirical bias corrections, monthly means of daily Tmax and Tmin values were analyzed. For each synthetic series, all daily temperature values were averaged by month and then compared to the corresponding statistics for the historical data.

At all 11 stations, the monthly means of the historical temperature data exhibit a similar annual cycle. The coldest temperatures occurred in January, and the warmest temperatures were observed in summer months. The warmest station is Miami (FL), where the warmest monthly mean occurred in August and was approximately 90 °F for Tmax and 77 °F for Tmin. The lowest monthly mean occurred at Blairsville (GA). Here, in January, the Tmax monthly mean was less than 50 °F and for Tmin was near 25 °F. The amplitudes of the annual cycle for different stations ranged from 15 to 40 °F. Typically, these amplitudes increased towards the northern part of the study area. In southern Florida, the amplitudes ranged from 15 to 20 °F, while at stations in northern Alabama and Georgia the amplitudes were near 40 °F.

After empirical temperature corrections were incorporated into the synthetic series, the weather generator successfully replicated historical temperature means. Representative results for monthly means from Atlanta (GA) are shown in Figure 3.1. The boxplots indicate the dispersion of the monthly means of the 100 synthetic series. In Atlanta (GA), Tmax means ranged from near 90 °F in June to almost 50 °F in January. For Tmin, the mean was near 70 °F in July and less than 35 °F in January. The historical
values coincided with the medians in all months for both Tmax and Tmin, and such a result was consistently observed across all stations.

3.1.2.2 Standard Deviation of Tmax and Tmin

The dispersion of Tmax and Tmin daily values is another diagnostic that can be used to assess the ability of the generator to simulate realistic weather. Some weather generators have been shown to underestimate the dispersion of daily Tmax and Tmin
(Wallis and Griffiths 1995; Hayhoe 1998). This statistic is important in many sectors; for example, changes in variability of temperature can impact crop growth predictions (Riha et al. 1996). For the historical data, a pattern consisting of the largest standard deviations in winter months and smallest values in the summer months appeared at all stations.

The annual cycles of the standard deviation of daily Tmax and Tmin were well replicated by the weather generator. For Atlanta (GA) in Figure 3.2, the standard deviation values for Tmax ranged from near 11 °F in January to approximately 5 °F in August; while for Tmin, the range of values was from near 11 °F in January down to approximately 3 °F in July. In Figure 3.2, the historical values tended to fall above the medians of the boxplots, but within the contiguous ranges, indicating slight underestimation of the dispersion of daily temperature values. However the deviations from the medians of the synthetic results (typically less than 0.5 °F) were relatively small. This result was typical for all other stations, where the historical values tended to lie within the contiguous ranges of the corresponding boxplots and relatively close to the median synthetic value.

Other generators have successfully replicated the annual cycle of this statistic including the model from Yates et al. (2003). They used a nonparametric k-NN generator at locations in the Midwestern United States, and found that the annual cycle of standard deviation of daily mean temperature (instead of Tmax and Tmin) was well captured by their model. At locations in Colorado and Illinois, monthly boxplots of standard deviation of daily temperature followed the pattern of the historical data with the highest historical values observed in the winter months and the lowest values observed in summer months.
In most cases the historical value fell within the interquartile range of the boxplots. Overall the graphic diagnostics in Yates et al. (2003) reveal that their k-NN model performed similarly to this model in reproducing this statistic.

3.1.2.3 Lag-1 Correlation for Tmax and Tmin

The persistence of daily Tmax and for daily Tmin values was diagnosed by calculating lag-1 correlation, defined as the relationship of values of a time series with values of that same time series on the previous day. Lag-1 correlation was calculated...
separately for wet and dry days. This was done because the initial calculations of lag-1 correlation for all days revealed a noticeable undersimulation of this statistic. Separate calculations for wet and dry days allowed for a more detailed diagnosis of lag-1 correlation in the synthetic data.

A temporal pattern for historical lag-1 correlation (when calculated for all days and for dry days only) was observed at the more southern locations in the study area. Here, coefficients were typically lower in summer months, especially for Tmin. This trend was more prominent at the two inland Florida peninsula stations: Clermont and Moore Haven. A possible contributor to this trend is the decrease in frequency of synoptic fronts (and accompanying air masses) during the summer in the study area. This is especially true in southern Florida, where Clermont and Moore Haven are located (Henry et al. 1994). For example, the incursion of cold air following cold fronts can influence temperature for several continuous days, holding Tmax and Tmin below their respective climatological means. This would contribute to a stronger persistence of temperature values in these months. The decreased occurrence of frontal systems in summer may correspond to decreased persistence.

Across the 11 stations, the annual patterns of dry day lag-1 correlations for dry synthetic days were captured; however, in most months, this statistic was underestimated, especially in summer months. Representative results for lag-1 correlation for dry days at Clermont (FL) are shown in Figure 3.3. A sharp drop off in lag-1 correlation in summer months can be observed. Typically, the historical values fell above the interquartile range of the synthetic lag-1 values. In non-summer months, the differences between the historical lag-1 coefficients and corresponding median synthetic values were small,
typically less than 0.05; however, in summer months, these differences were larger (over 0.1 for Tmin in June). The patterns observed for this diagnostic at Clermont (FL) are typical of what was seen at other stations.

At all locations, monthly wet day lag-1 correlations of synthetic data were consistently lower than the historical values. For this statistic, there was no apparent annual pattern across stations. Diagnostics for wet day lag-1 correlations at Clermont (FL) are shown in Figure 3.4. The correlation coefficients for Tmax and Tmin were underestimated in all months. In most months, the historical values fell outside of the
Figure 3.4: Boxplots of wet day lag-1 correlation of a) daily Tmax (°F) and b) daily Tmin (°F) at Clermont (FL) for each synthetic series, calculated separately for each month of the year. The boxplots indicate the dispersion of the statistics for the synthetic series. The circles joined by lines indicate the corresponding statistics for the historical series.

The contiguous range of the boxplots; the median values of the synthetic wet day correlations were typically 0.2 to 0.3 lower than the historical values.

Generally, it has been claimed that lag-1 correlations are well reproduced by the k-NN model (Rajagopalan and Lall 1999). For wet day lag-1 correlation, one possible explanation for the underestimation could be that the separation of wet days into two states would reduce the ability of the k-NN selection procedure to capture the persistence of temperature values on wet days. Warming and cooling trends consist of daily Tmax and Tmin steadily increasing or decreasing over a period of days and these trends would
contribute to persistence of temperature values in the historical data. The transitioning between wet and very wet days could disrupt the simulation of warming or cooling trends during wet spells by the generator. During transitions, the number of eligible neighbors is limited by the availability of appropriate pairs. The number of pairs available is typically small when transitioning from a wet to very wet day (or vice versa). With a small number of available neighbors, it is likely that the number of days with similar Tmax and Tmin values to that of the current synthetic day would be limited or non-existent. Thus, it is likely that the successor to the current synthetic day will have substantially different temperature values, and any cooling or warming trend in the synthetic series would stop. Conversely, these disruptions would not occur in wet spell simulations in a non-parametric k-NN model with two states.

This hypothesis was explored in detail at Clermont (FL) and it was found that for transitions (wet to very wet or very wet to wet), lag-1 correlation was more strongly underestimated than for consecutive wet days only (wet followed by wet days). However, there was still noticeable underestimation for lag-1 correlation on wet day pairs. Since underestimation was greater on transitioning days during wet spells, this analysis was taken a step further and transitions between dry and wet days were analyzed. It was found at Clermont (FL) that lag-1 correlations for days transitioning between dry and wet/very wet states (dry to wet, wet to dry, dry to very wet, etc.) were relatively more underestimated than lag-1 correlations for dry days (dry days following dry days).

3.1.2.4 Skewness of Tmax and Tmin

Skewness coefficients for distributions of daily Tmax and Tmin values were estimated by month from the synthetic sequences and the historical data. Skewness is
defined as a measure of the asymmetry of a distribution. A skewness coefficient of zero indicates that the distribution is perfectly symmetric. A positive coefficient indicates that the distribution is skewed to the right, which means there is a longer right tail (e.g. a gamma distribution). The opposite is true for a negative skewness coefficient.

There were similarities across all stations in the temporal patterns of skewness of Tmax and Tmin in the historical data. In most months, skewness coefficients for both Tmax and Tmin were negative. Harmel et al. (2002) attributed the negative skewness of temperature values in the SEUS to the occasional passage of cold fronts (followed by cold air masses) and to cloudy days with precipitation, two phenomena that often prevent daily temperatures from reaching their climatological mean value. An additional pattern observed at most locations was that skewness coefficients were more negative in the summer months as compared to other times of the year. This may be due to the decreased number of cold spells and the longer duration of warm spells associated with summers in the southern United States (Henderson and Muller 1997).

The temporal patterns of skewness coefficients were well captured by the generator for both Tmax and Tmin at all locations. The skewness results for Mobile (AL) for this work are shown in Figure 3.5. For both Tmax and Tmin, the historical skewness coefficients fell within the contiguous range of boxplots of synthetic values in all months. This was typical for all other locations. For Tmin at a few locations in a few months, the historical values were not within the contiguous ranges of respective synthetic values. However, the distributions of historical data are often irregular and outliers or other irregular characteristics (bimodality, etc.) could significantly impact a skewness estimate.
Thus, the occurrence of deviations in only a few cases was not considered of practical significance.

The historical characteristics of skewness observed here were consistent with research by Harmel et al. (2002). They analyzed monthly distributions (for January, April, July, and October) of daily Tmax and Tmin in several U.S. cities in the SEUS, and found that distributions of these variables were often negatively skewed. For example, at Mobile (AL) they found that distributions of Tmax were negatively skewed in January, April, July and October, and were negatively skewed for Tmin in April, July, and
October (but positively skewed in January). This matches what was observed in this work for Mobile (AL) in Figure 3.5.

3.1.2.5 Cross Correlation of Tmax and Tmin

Cross correlation is a diagnostic used to measure the relationship between two time series of different variables (here, Tmin and Tmax). Calculations were done separately for wet and dry days because of differences in the magnitude of cross correlation between precipitation states. Typically, historical monthly wet day coefficients were lower than corresponding dry day coefficients by a magnitude of 0.1 to 0.4.

The observed temporal patterns for both wet and dry day cross correlation were similar for all stations, with the lowest values occurring in summer months. The seasonal variations in this statistic may be related to changes in phenomena which influence weather across seasons. In summer months convective influences are stronger (Soule 1998), and heating and cooling effects upon temperature due to convection may change from day to night. For example, an afternoon rain shower may impact Tmax, but any cloud cover may dissipate before nightfall when Tmin usually occurs; thus, what affects Tmax may not affect its corresponding Tmin. In non-summer months, the behavior of Tmax and Tmin is more strongly impacted by fronts (Soule 1998). Effects from these systems persist for several days and thus influence both Tmax and the corresponding Tmin.

Cross correlation of Tmax and Tmin was well replicated by the weather generator on both dry days and wet days. The cross correlation diagnostic for dry days at Atlanta (GA) is shown in Figure 3.6 and for wet days in shown in Figure 3.7. In both plots, the
Figure 3.6: Boxplots of dry day cross correlation coefficients for daily Tmax and daily Tmin at Atlanta (GA) for each synthetic series, calculated separately for each month of the year. The boxplots indicate the dispersion of the statistics for the synthetic series. The circles joined by lines indicate the corresponding statistics for the historical series.

Historical correlation coefficients and corresponding median synthetic values (indicated by the horizontal lines within the boxes) were lowest in July. In most months, the historical value coincided with the median of the boxplots indicating that the historical statistics were a likely realization of the generation process. Results shown here are representative of the diagnostics at all other locations.

Rajagopalan and Lall (1999) showed that their k-NN generator successfully simulated this statistic. They compared the performance of their k-NN generator to that of a version of the parametric model by Richardson (1981) and found that the k-NN model
Figure 3.7: Boxplots of wet day cross correlation coefficients for daily Tmax and daily Tmin at Atlanta (GA) for each synthetic series, calculated separately for each month of the year. The boxplots indicate the dispersion of the statistics for the synthetic series. The circles joined by lines indicate the corresponding statistics for the historical series.

reproduced cross correlation more accurately. They noted that the k-NN process should reproduce cross correlation well, as all values of all weather variables for a given day (in the historical data) are selected together by the nearest neighbor procedure.

3.1.2.6 Dates of First Freeze and Last Freeze (Indicators of Frost Free Period)

The average dates of last freeze (DLF) (Tmin ≤ 32 °F) and first freeze (DFF) were estimated from the synthetic series. The difference between these two values indicates the number of days in the frost free period (FFP). Estimates of FFP are important for
agriculture and are used in assessing growing season length and predicting crop
production (Hayhoe and Stewart 1996).

For the 11 stations, the calendar day of the average historical DFF and DLF
varied among locations. At Miami (FL), there were no freezing days in some years. In the
years in which there were freezing days, the historical mean for DLF was January 16th
and for DFF was January 22nd. Conversely, for Blairsville (GA) the average historical
value for DFF was October 7th (day 280 in a non-leap year) and for DLF was May 1st (in
a non-leap year day 121). The average FFP was calculated as the difference between
these two dates and at Blairsville (GA) this equals 159 days.

Values of DLF, DFF, and FFP were well replicated by the weather generator. It
was hypothesized that since the k-NN procedure uses only the historical data within a
limited temporal window, DLF and DFF should be well replicated. However, the use of
empirical bias corrections could have had an adverse impact. Representative results for
this diagnostic are shown for Atlanta (GA) in Figure 3.8, a histogram of average DLF and
DFF. For DLF (first panel), the historical value was day 82 (March 23rd in a non-leap
year) while the mean DLF of the synthetic series was day 83 for a difference of one day.
For DFF (second panel), the historical value fell on day 312 (November 8th in a non-leap
year) while the mean of the synthetic values also fell near day 312. The average historical
FFP at this station was 230 days, and the corresponding value for the synthetic data was
229 days – a difference of one day. Across all other locations, the average DLF and DFF
was no more than five days different than the corresponding historical values. The
deviation in synthetic FFP ranged from seven days shorter (at two locations) to one day
longer (at one location) than the corresponding historical value across the 11 locations.
Figure 3.8: Histograms of the average a) day of a last freeze (Tmin <= 32 °F) and b) day of first freeze, calculated for each synthetic series at Atlanta (GA). The grey lines indicate the corresponding values for the historical series.

This model performed better than the parametric models studied by Hayhoe and Stewart (1996) for FFP replication. Hayhoe and Stewart (1996) compared the performance of two parametric generators using data from stations in Canada. They found that both models underestimated average FFP at all locations tested; the difference between synthetic and historical values ranged from six days to 23 days.

3.1.2.7 Hot and Cold Spells

The replication of continuous sequences of very hot or very cold days (hot spells and cold spells) was analyzed to determine how well the model generated extreme temperatures. Hot and cold spells can have significant impacts on some human sectors.
For example, increased mortality rates have been associated with increased occurrence of consecutive very hot days (Kalkstein et al. 2008). Hot spells also can affect energy demand and decrease crop growth and yield (Mearns et al. 1984). To assess this generators’ reproduction of hot spells and cold spells, the number of continuous runs (five days or longer) of hot or cold days was studied. Hot days were defined as having a Tmax above the 0.90 quantile of the entire Tmax record and cold days were defined as having a Tmin below the 0.10 quantile of the entire Tmin record. The numbers of hot and cold spells were summed over the entire historical and synthetic sequences.

The diagnostics revealed some similarities among the stations for historical spell statistics. For most stations, the total number of hot spells ranged from 60 to 100 and the number of cold spells ranged from 45 to 70 over the entire historical sequences (from 57 to 60 years). At all locations except Miami (FL), the number of hot spells exceeded the number of cold spells. In the SEUS, extremely high temperatures are often related to the advection of maritime tropical air from the Gulf of Mexico in the summer; and the persistence of high temperatures has been associated with the strength of the Atlantic subtropical high (Henderson and Muller 1997). Cold spells are often due to polar air masses associated with anticyclones in the winter (Henderson and Muller 1997).

There were large differences between synthetic and historical values for these statistics at some locations, however it should be noted that this diagnostic is a summation across the entire sequences (which range from 57 to 60 years). An example of this diagnostic is provided for Mobile (AL) in Figure 3.9. For hot spells, the historical value represented by the vertical line was below the central tendency of the histogram. The mean synthetic value was 107, while the historical value was 90 spells (for a
Figure 3.9: Histogram of synthetic a) hot spells and 2) cold spells at Mobile (AL). The histograms are of totals of each spell type over each synthetic series. The grey lines indicate the corresponding totals over the historical series.

difference of 17 spells). However, recall that this plot was a summation over the entire series, and the data at Mobile (AL) was 59 years in length. This would mean that the number of hot spells was underestimated by less than one every three years (17 spells over 59 years). For cold spells at this station (right panel) the historical value fell near the mean synthetic value indicating that this statistic was well replicated.

There was some variability in the results among different locations when comparing the average synthetic values (of total hot and cold spells) with the historical values. The total number of hot spells was overestimated by 10 or more at five locations (the largest difference was 21 at Miami, FL) and underestimated by 10 at one location.
The total number of cold spells was overestimated by as much as eight or underestimated by up to seven. Conversely, at five locations the total number of synthetic hot spells was within one of the historical value; and at five locations (not the necessarily the same) the total number of synthetic cold spells was within three of the historical value.

3.1.3 Assessment of Precipitation Generation

The ability of the generator to reproduce the characteristics of daily precipitation (occurrence and intensity) and statistics covering longer time spans such as annual totals was tested using graphic diagnostics. The performance of both the Markov chain and the k-NN process were assessed. In the semi-parametric generator, days were categorized as “dry”, “wet”, or “very wet” depending on the daily precipitation amount (see section 2.3.1). For the purpose of these diagnostics, both “wet” and “very wet” days were labeled as wet (except where noted). In addition to assessing the models performance, the climatology of precipitation in the SEUS was characterized through examination of the temporal patterns of precipitation statistics across the 11 locations.

3.1.3.1 Probability of a Wet Day

The probability of a day being wet or very wet is one means of assessing the precipitation occurrence process. Wet day probabilities were calculated separately for each month and synthetic set. For this diagnostic, the historical data exhibits some common geographic patterns. For many stations, the lowest precipitation probabilities occurred in late spring (April or May) or fall (October); and the highest probabilities appeared in summer months. At the three stations in southern Florida (Miami, Moore Haven, and Clermont), there were two different periods in the annual patterns. A “dry”
(relatively lower probabilities) period encompassed the months from November to April, and a “wet” (relatively higher probabilities) period existed from June to September.

The temporal patterns observed in these diagnostics correspond with the climatological behavior of precipitation in the SEUS. The dominant form of precipitation in the SEUS from November to March is related to frontal activity, however many fronts do not reach the southern part of Florida due to a high pressure system in the Atlantic Ocean (Henry et al. 1994). In addition to blocking the passage of fronts, this high pressure system also inhibits convective activity during winter in southern Florida. This high pressure system is likely the reason for the low rainfall probabilities observed in southern Florida from November to April. Alternatively, the dominant form of precipitation in the study area during the late spring and summer is convective (Henry et al. 1994); and, combined with hurricane influences, are likely the reasons for the high wet day probabilities seen in the summer in southern Florida. The lower wet day probabilities observed in April, May, and October at some locations likely correspond to the lull between the strongest frontal and convective activity in the region (Henry et al. 1994).

Overall, diagnostics revealed that the Markov chain does an excellent job of modeling wet day occurrence. Representative results for this statistic are shown for Atlanta (GA) in Figure 3.10. In all months, the historical value coincided with the median of the boxplots indicating that the historical statistics were a likely realization of the generation process. This type of result was seen at all locations. These results agree with previous work which found that first-order Markov chains were appropriate for generating daily precipitation occurrence at stations in the Eastern United States (Wilks 1999).
Figure 3.10: Boxplots of the probability of a wet day (precipitation ≥ 1.2 HI) for each synthetic series, calculated separately for each month of the year at Atlanta (GA). The boxplots indicate the dispersion of the statistics for the synthetic series. The circles joined by lines indicate the corresponding statistics for the historical series.

3.1.3.2 Probability of a Very Wet Day

Probabilities of very wet days were well modeled by the Markov chain. Wet and very wet days were separated in the Markov chain by the 80th percentile of the monthly distribution of daily precipitation. These values are shown in Table 3.1 and ranged from a minimum of 47 HI (Moore Haven, FL in December) to a maximum of 122 HI (Mobile, AL in March). The 80th percentile values for the annual distributions of daily precipitation are also shown. No diagnostics were shown for this statistic, but for all 11
stations, the historical value coincided with the median of the synthetic values in all months. The temporal patterns of the historical data observed for this statistic matched those observed for wet day probabilities.
3.1.3.3 Mean Wet and Dry Spell Lengths

Another diagnostic used to check for accurate reproduction of precipitation occurrence was the assessment of the duration of spells of consecutive wet and dry days. Statistics of this metric (such as mean, median, and 90\textsuperscript{th} percentile) were compared with corresponding values for the historical data. This metric is important in various decision models such as agricultural yield, watershed, and water reservoir models (Yates et al. 2003). For example, in agricultural yield models, dry spell occurrence affects crop water availability, which in turn influences crop yields (Barron et al. 2003).

Analyses of the historical mean spell lengths revealed different patterns among locations. The longest mean wet spell lengths were in the summer for most stations. For the three Florida peninsula stations (Miami, Moore Haven, and Clermont) there were two different periods for mean dry spell length. The summer months have the shortest mean dry spell lengths (fluctuating around two days), while mean dry spell lengths in non-summer months fluctuate between six and eight days. The patterns observed at all stations correspond with patterns observed with wet day probabilities (i.e. maximum dry spell lengths occurring the same time as smallest wet day probabilities).

Mean spell lengths for dry and wet days were well replicated across all stations. Representative results of the mean spell length statistics are shown for Miami (FL) in Figure 3.11. This plot illustrates the strong difference between summer and winter months for spell length statistics in the Florida peninsula. In the first panel (wet day spells), the monthly means of historical wet spell lengths from November to April were around 1.5 days, while the means from June to September were near 2.5 days. For dry spells, the monthly mean lengths were between 2 and 3 days in the summer months and
near or greater than 6 days in winter months. For all months, the historical mean lies near the corresponding median of synthetic mean lengths (indicated by the horizontal line inside the boxplots).

3.1.3.4 Longest Annual Wet and Dry Spell Lengths

To check for the replication of extreme spell lengths, the longest wet and dry spell lengths were determined for each year in each synthetic sequence and the historical data. This statistic is important, for example, because it can be used as an indicator of drought. There are some geographic trends in the SEUS for this quantity. At most stations, the median historical values of maximum annual dry spell lengths were near 20 days. At
southern Florida stations, the medians of maximum annual dry spell lengths were a little higher, near 25 days. At the two northernmost stations, Belle Mina (AL) and Blairsville (GA), the median lengths were closer to 15 days. For annual longest wet spells, the median historical values were six or seven days at most locations.

The weather generator does a satisfactory job for this diagnostic. Representative results are provided for Atlanta (GA) in Figure 3.12. In this figure, the boxplots show the dispersion of the maximum annual spell lengths for the historical data (H) and for comparison purposes, the first four synthetic series (S1, S2, S3, and S4). For dry spell lengths (first panel), the medians of the values for the synthetic series were similar to the median historical value. In addition, the interquartile ranges were relatively similar across all boxplots indicating that the dispersion of central values was well reproduced. There was some variation in the contiguous ranges of boxplots, but the different ranges for the synthetic values fluctuate close to the range of the historical data. For wet spell lengths, the medians were identical across the historical data and the first four synthetic series. In addition, the interquartile ranges of boxplots were all identical, while the contiguous ranges were all similar. These types of results were consistent across most stations. The median maximum dry spell and wet spell lengths of the synthetic series were typically within one day of the corresponding historical value.

Wilks (1999) studied this statistic (for months instead of years) at stations and found that first-order Markov chains were “adequate” for simulating extreme spell lengths at locations in the eastern United States.
3.1.3.5 Daily Precipitation Intensity

Accurate reproduction of distributions of daily precipitation amounts is important for soil erosion studies and flood risk assessment (Semenov et al. 1998). To check for the accurate generation of distributions of daily precipitation amounts, quantile-quantile (Q-Q) plots were used. A Q-Q Plot (Figure 3.13) is a graphical tool used to illustrate the differences between two distributions by comparing a range of estimates in the data sets. If the distributions are similar, then points should fall near the 1:1 line (slope = 1). If points were above the 1:1 line, then the distribution of values represented by the y-axis

Figure 3.12: Boxplots of the duration of the longest a) dry spells and b) wet spells in each year at Atlanta (GA). The boxplots indicate the dispersion of spell lengths (one per year) over the historical series (H) and the first four synthetic series (S1, S2, S3, S4).
(in this case, the simulated sequences) is displaced to the right of the distribution represented by the x-axis (the historical data). The reverse is true if points are below the 1:1 line.

Distributions of daily precipitation were generally well simulated; however, diagnostics revealed some discrepancies for high daily precipitation amounts at some locations. Distributions of daily precipitation were well replicated in Miami (FL) (Figure 3.13). For this plot, all symbols fell near the 1:1 line indicating that the distributions were similar. For this diagnostic at Mobile (AL) (Figure 3.14), some of the upper values fell beneath the 1:1 line. At stations where there was a noticeable deviation from the 1:1 line, the differences consistently occurred at values above 200 HI (or two inches). It was
hypothesized that the presence of a few extremely high values in the historical data sets may have caused this deviation. To test this hypothesis at Mobile (AL), a new Q-Q plot (not shown) was created where the two largest historical values were removed from both the historical and synthetic data. In this plot, the deviation from the 1:1 line at higher values decreased. For Mobile (AL), some of the deviation from the 1:1 line was due to the presence of a few large outliers in the historical data; and this may be the case at other stations where deviations occurred.

3.1.3.6 Monthly Total Precipitation

An aggregate precipitation diagnostic was the median of monthly precipitation totals. This diagnostic illustrates the combined performance of both the occurrence and
intensity processes. For this statistic, precipitation was summed for each month and each synthetic series. The medians of the monthly totals for each synthetic series were calculated and then compared to the corresponding historical values. For this statistic, there were a few geographic trends across stations in the historical data. In southern Florida, the medians of monthly totals were highest in summer months and lowest from December through March. A second annual pattern observed was apparent at stations in middle and upper Alabama and Georgia. Here, the highest median totals occurred in March and the lowest totals occurred in October. These patterns generally match the characteristics of climate of the SEUS.

Diagnostics reveal that the temporal patterns of the historical precipitation totals were captured by the generator. At Tifton (GA) (Figure 3.15), in all months, the historical median values fell within the contiguous range of the boxplots of synthetic values; and the temporal pattern of the boxplots matched the temporal trend of the historical values. This type of result is typical of other locations.

Gangopadhyay et al. (2005) tested a k-NN model on station data in south-central Georgia and among the diagnostics they used were the annual cycle of total monthly precipitation. The temporal pattern for this statistic observed by Gangopadhyay et al. (2005) was similar to that observed in Figure 3.15 (Tifton, in southern Georgia). Their model replicated the annual cycle well. In all months, the historical values fell within the contiguous ranges of the boxplots. For both models in this region, the k-NN procedure satisfactorily captured this precipitation statistic.
3.1.3.7 Annual Total Precipitation

Annual totals also were calculated as an aggregate precipitation diagnostic. For each synthetic sequence and the historical data, the mean annual total was determined. In Table 3.2, the mean annual totals of the historical data and corresponding medians of the mean annual totals for the synthetic data are shown for each station. The table is sorted by historical value with the lowest values coming first. No geographic patterns were apparent; however the two coastal stations have the highest historical totals. The largest difference between synthetic totals and corresponding historical totals was 100 HI at Chipley (FL).

Figure 3.15: Boxplots of the median monthly total precipitation (HI) (for wet days only) for each synthetic series, calculated separately for each month of the year at Tifton (GA). The boxplots indicate the dispersion of the statistics for the synthetic series. The circles joined by lines indicate the corresponding statistics for the historical series.
<table>
<thead>
<tr>
<th>Station</th>
<th>Historical Means (HI)</th>
<th>Median Value of Synthetic Means (HI)</th>
<th>Difference (HI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brooklet</td>
<td>4639</td>
<td>4595</td>
<td>44</td>
</tr>
<tr>
<td>Tifton</td>
<td>4656</td>
<td>4566</td>
<td>90</td>
</tr>
<tr>
<td>Moore Haven</td>
<td>4778</td>
<td>4730</td>
<td>48</td>
</tr>
<tr>
<td>Atlanta</td>
<td>4921</td>
<td>4891</td>
<td>29</td>
</tr>
<tr>
<td>Clermont</td>
<td>4991</td>
<td>4894</td>
<td>97</td>
</tr>
<tr>
<td>Montgomery</td>
<td>5076</td>
<td>5032</td>
<td>44</td>
</tr>
<tr>
<td>Belle Mina</td>
<td>5283</td>
<td>5208</td>
<td>75</td>
</tr>
<tr>
<td>Chipley</td>
<td>5542</td>
<td>5442</td>
<td>100</td>
</tr>
<tr>
<td>Blairsville</td>
<td>5602</td>
<td>5545</td>
<td>57</td>
</tr>
<tr>
<td>Miami</td>
<td>5977</td>
<td>5969</td>
<td>8</td>
</tr>
<tr>
<td>Mobile</td>
<td>6560</td>
<td>6505</td>
<td>55</td>
</tr>
</tbody>
</table>

Table 3.2: Annual precipitation totals at the 11 stations. The second column is the historical mean. The third column is the median of the means of the 100 synthetic series. The fourth column is the difference between the values in the second and third columns.

The generator satisfactorily models this statistic, although it should be noted that the synthetic results were generally slightly lower than the historical results. A histogram of synthetic means annual totals for Atlanta (GA) is shown in Figure 3.16. The historical value is indicated by the vertical line and fell within the histogram, but not close to the central tendency of simulated values. Nevertheless, the difference between the historical value and the median of the synthetic values was 29 HI, which was a small percentage (0.6%) of the historical annual total (4921 HI). At most locations, the historical values fell in the upper half of the corresponding histograms. The exception to this was at Miami (FL) where the historical value fell near the center of the histogram.
3.1.3.8 Standard Deviation of Daily Precipitation

The dispersion of daily precipitation totals is an important statistic that has not always been well replicated by weather generators. Under-simulation of this statistic can be one of the factors that contributes to “overdispersion”, or the tendency of weather generators to underestimate interannual variability in monthly (or seasonal) total precipitation (Wilks 1989). In this diagnostic, the standard deviation of daily precipitation values for each month was estimated. The only consistent trend in the historical data observed was that the historical values in the spring and/or fall months tended to be higher as compared to other times of the year.
The annual cycles for this statistic were captured by the generator at all locations. For Atlanta (GA) (Figure 3.17), the historical values fell within the interquartile ranges of the boxplots in most months, and otherwise fell within the upper or lower quartile. This result is consistent across all locations.

Wilks (1999) studied this statistic (using Q-Q plots) at stations in the United States and found that some parametric methods tended to underestimate the standard deviation of monthly distributions of daily precipitation amounts. He found that the gamma distribution under-simulated standard deviations at most stations tested, while the mixed exponential distribution performed much better, although still under simulating standard deviation at some locations. Yates et al. (2003) used this same diagnostic with their k-NN generator. They found that their weather generator satisfactorily captured the temporal pattern of this statistic at two different locations, with the historical values falling within the contiguous range of the boxplots in most months. The results in this work indicated that this model performs similarly to the k-NN model by Yates et al. (2003) and better than the parametric methods from Wilks (1999) for this diagnostic.

3.1.3.9 Standard Deviation of Monthly Total Precipitation

As mentioned previously, the interannual dispersion of monthly (or seasonal) total precipitation is often not well replicated by weather generators (e.g. Katz and Parlange 1998; Wilks 1999; Mason 2004). This can, for example, lead to distortion in the variability of predicted yields from crop model simulations (Mavromatis and Hansen 2001). These synthetic series were tested for overdispersion by calculating the interannual standard deviation of total monthly precipitation. There were no significant temporal or geographic patterns among the stations for this statistic.
The weather generator captures the annual pattern of this diagnostic; however, there were large differences between historical and median synthetic statistics in a few months at some locations. This diagnostic for Brooklet (GA) is provided in Figure 3.18. In all months, the historical statistics were a possible realization of the generator process lying within the contiguous ranges of the boxplots. In some months, the statistic was well simulated with the historical values falling close to the median and within the interquartile range. However, in a few months there were prominent differences between the median synthetic and corresponding historical statistics. For example, in June, the historical value was over 50 HI greater than the corresponding synthetic median value.

Figure 3.17: Boxplots of the monthly standard deviation of daily precipitation (HI) for each synthetic series, calculated separately for each month of the year at Atlanta (GA). The boxplots indicate the dispersion of the statistics for the synthetic series. The circles joined by lines indicate the corresponding statistics for the historical series.
This type of result was seen at other stations with some variability in capturing historical statistics among months and locations. Underestimation of standard deviations occurred more often in summer and early fall months. However, in a few months, there was overestimation of the dispersion of total monthly precipitation with the historical statistics falling in the lower quartile of the boxplots of synthetic values.

3.2 Synthetic Series Conditioned on Climate Forecasts

Different approaches have been used for creating synthetic series conditioned on climate information through the use of weather generators. These approaches can be classified into three groups: (1) conditioning the parameters of parametric weather
generators (e.g. Wilks 2002), (2) resampling the input (i.e. historical data) for weather
generators based on climate information (e.g. Apipattanavis et al. 2007), and (3)
modifying the synthetic output to match target statistics (e.g. Hansen and Ines 2005).
Here, a new approach was used where the synthetic output of a weather generator was
resampled to produce distributions of output values consistent with a particular forecast.

In this work, the output of the semi-parametric generator was resampled based on
seasonal climate forecasts in the SEUS in two different formats. One format was the
traditionally used tercile forecast format (format I). Resampling based on forecasts in this
format requires knowledge of site-specific boundaries between terciles (corresponding to
the 0.33 and 0.66 quantiles of historical distributions of precipitation and temperature).
The second format (format II) utilized in this work includes the entire forecast
distribution.

In this section, first, the estimates of the tercile boundaries by different methods
were compared and the causes of differences among methods were explored. Then, the
synthetic output from the resampling approach was analyzed using graphic diagnostics.
The conditioned sets of synthetic series were examined to assess if they reflected the
Corresponding forecast scenarios and the characteristics of daily precipitation were
studied.

3.2.1 Assessment of Tercile Boundary Estimates

In order to produce synthetic series conditioned on tercile format climate
forecasts, the numerical boundaries separating terciles were needed for each location
studied. The boundaries separating terciles – or the 0.33 and 0.66 quantiles (Q33 and
Q66) – were estimated at all cooperative stations in the SEUS using four different
methods for quarterly average daily temperature (Tavg) and quarterly total precipitation. Diagnostics were used to analyze the estimates produced by different methods. In most cases, all methods produced similar estimates; the data sets where quantile estimates differed were studied in detail. In addition, the variability in quantile calculation introduced by incomplete data records was explored by estimating Q33 and Q66 from data records of different lengths.

3.2.1.1 Q33 and Q66 Estimates of Temperature and Precipitation

All four methods typically produced similar estimates and spatial patterns for the target quantiles, thus only results for the empirical (EMP) approach are displayed and discussed. Estimated EMP Q33 (Q66) values for Tavg range over the entire study area from 4.24°C (5.56°C) to 28.73°C (29.11°C). The range for precipitation Q33 (Q66) values goes from 81 mm (136 mm) to 636 mm (770 mm). The minimum values occurred in JFM and the maximum values were found in JAS for both variables. Figures 3.19 through 3.22 show contour maps of EMP Q33 and Q66 values for Tavg and precipitation. The spatial patterns for quantiles of both variables were, as expected, similar to plots of quarterly climatological Tavg and precipitation values available from the Climate Diagnostics Center (CDC, see http://www.cdc.noaa.gov). Precipitation patterns were more variable among quarters than corresponding Tavg contours, and the largest gradients tended to be found in coastal regions.
Figure 3.19: Contour plots of empirically-derived Q33 estimates for quarterly mean surface temperature (°C). The four panels illustrate values for each quarter of an “ENSO year” in the southeastern United States. They are arranged clockwise from top left: October-December (OND), January-March (JFM), April-June (AMJ) and July-September (JAS). Station locations are shown in the lower right-hand panel of Figure 3.19.
Figure 3.20: Contour plots of empirically-derived Q66 estimates for quarterly mean surface temperature (°C). The four panels illustrate values for each quarter of an “ENSO year” in the southeastern United States. They are arranged clockwise from top left: October-December (OND), January-March (JFM), April-June (AMJ) and July-September (JAS). Station locations are shown in the lower right-hand panel of Figure 3.19.
Figure 3.21: Contour plots of empirically-derived Q33 estimates for quarterly total precipitation (mm). The four panels illustrate values for each quarter of an “ENSO year” in the southeastern United States. They are arranged clockwise from top left: October-December (OND), January-March (JFM), April-June (AMJ) and July-September (JAS). Station locations are shown in the lower right-hand panel of Figure 3.19.
Figure 3.22: Contour plots of empirically-derived Q33 estimates for quarterly total precipitation (mm). The four panels illustrate values for each quarter of an “ENSO year” in the southeastern United States. They are arranged clockwise from top left: October-December (OND), January-March (JFM), April-June (AMJ) and July-September (JAS). Station locations are shown in the lower right-hand panel of Figure 3.19.
3.2.1.2 Divergence of Estimates by Different Methods of Q33 and Q66

The estimates of Q33 and Q66 from distributions of seasonal total precipitation totals and mean surface temperatures in the SEUS were compared through graphic diagnostics. Scatterplots comparing the estimates of all combinations of any two methods were plotted for both Q33 and Q66 for each quarter, method, and variable. The scatterplots revealed that most estimates for Q33 and Q66 by the different methods were similar.

As an example, Figure 3.23 shows a scatterplot matrix of all Q66 values for Precipitation in JFM. Each point in a panel represents this quantile for a given station, estimated by a pair of methods X and Y. The methods used in each panel are identified in the corresponding column for method X and the corresponding row for method Y. If quantile estimates coincide between two methods, then all points should fall near the 1:1 line in each panel. In Figure 3.23, all four methods generally produced similar results. However, in a few cases, estimates differed.

To explore what might cause estimates to diverge, the largest 10% of differences in quantile estimates between any two methods for each quantile and variable were examined. For each series considered, a histogram of the historical values was plotted and fits produced by each method were overlaid. These plots were analyzed to look for associations between characteristics of distributions and the divergence of estimates.

Many of the histograms of historical values for which estimates diverged were irregularly shaped. Irregular shapes included both negatively and positively skewed distributions, rectangular shape distributions, and distributions containing more than one mode. Other histograms for which estimates differed had outliers or gaps.
One common observation was the flexibility of the non-parametric methods, KERNEL and SPLINE, which can capture better irregularly shaped distributions. For example, the KERNEL method was able to capture outliers without altering its fit to the bulk of the distribution. On the other hand, these non-parametric methods can be...
hypersensitive and, in some cases, one or both of these methods produced fits that were too jagged or had more modes than the histogram.

Another finding was that the SPLINE method in some cases produced a poor fit to the data. In such cases, the SPLINE method was unable to capture the characteristics of the empirical distributions, and instead, produced a curve that was too smooth. Because of this, a second set of estimates was calculated by forcing the algorithm to use a predefined number of knots (as discussed in section 2.5.5). In some cases, the new estimates no longer showed large differences with estimates from other methods.

Some of the empirical distributions associated with large differences in quantile estimates between methods displayed bimodality, which could not be captured by all methods. One particular instance was inspected in further detail.

The historical series selected was JFM precipitation for Lisbon (FL). Figure 3.24 shows the histogram of this series and the fits by the different methods. Q66 estimates for this series by the GAMMA, KERNEL, EMP, and SPLINE methods were, respectively: 292, 316, 318, and 329 mm. Thus, the largest difference (37 mm) was between the GAMMA and SPLINE estimates. The apparent bimodality is not captured by the GAMMA fit, which puts too much probability density between the peaks. In contrast, both the KERNEL and SPLINE methods provide a better description of the bimodality.

What causes the apparent bimodality? Because precipitation and temperature in the SEUS are strongly influenced by the El Niño-Southern Oscillation (ENSO) phenomenon during the winter (La Niña winters tend to be warm and dry, while El Niño winters tend to be cool and wet) (Ropelewski and Halpert 1987), the potential impact of ENSO upon the historical data distribution was explored.
Each year in the historical record for Lisbon (FL) was assigned an ENSO phase (El Niño or warm events, La Niña or cold events, or neutral) following the definition by Florida State University’s Center for Oceanic and Atmospheric Prediction (COAPS, see http://www.coaps.fsu.edu). In Figure 3.25, histograms of precipitation JFM totals were plotted separately for each ENSO phase. Intermediate and high precipitation totals tended to occur more often in El Niño events. In contrast, during La Niña events, low precipitation totals were clearly more frequent. Neutral years had a fairly uniform
distribution of precipitation values. Thus, for this historical precipitation record, the bimodal shape could indicate a mixture of distributions associated with different extreme ENSO phases.

With instances of large differences with $T_{avg}$, many of the histograms for these data records did not display bimodality. However, since all $T_{avg}$ large differences occurred in OND and JFM, a possible link with ENSO was considered. When separate histograms were plotted by ENSO phase (such as those in Figure 3.25), lower values were more frequent during El Niño events and/or higher values occurred more often during La Niña events.

Figure 3.25: Histogram of precipitation (mm) historical values during JFM for Lisbon (FL) in (a) La Niña years (b) El Niño years and (c) Neutral years.
3.2.1.3 Impact of Missing Data upon Estimation of Q33 and Q66

Available historical data sets are often short and can have varying proportions of missing data. These characteristics can influence the variability of quantile estimates. In this section, series with varying simulated percentages of missing data were used to explore the impact of incomplete series on empirical quantile estimates.

First, the most complete historical precipitation series was selected: the record for Miami (FL) had 64 available OND values. Next, incomplete data sets of various sizes (20, 30, 40, and 50) were simulated by sampling without replacement from the full historical series. These incomplete data sets had percentages of missing quarterly values of 69%, 53%, 38%, and 22% respectively. For each record size, 1000 different simulated sets were generated. Then, Q33 and Q66 were estimated for each simulated data set using the empirical method (no bootstrapping).

Figure 3.26 shows boxplots of quantile estimates for each simulated sample size. It is apparent that the range of empirical estimates rose as the percentage of missing data increased. Overall, these results indicated that the accuracy of empirical quantile estimates decreases with higher proportions of missing data.

3.2.2 Assessment of Precipitation Values from Conditioned Synthetic Series

The resampling approach was tested using output from the weather generator by Apipattanavis et al. (2007). Synthetic output from two stations, Chipley (FL) and Miami (FL) was resampled based on hypothetical seasonal climate forecasts to create conditioned sets of 50 synthetic series. Chipley (FL) and Miami (FL) were selected because of the stronger ENSO teleconnections found in Florida as compared to Northern
Figure 3.26: Boxplots of empirical (not bootstrapped) estimates of Q33 and Q66 from simulated incomplete data sets of various sample sizes (20, 30, 40, and 50 values). The horizontal dashed line is the empirical estimate from the entire data set for a given quantile. Note that the y-axis changes for each quantile. The data used is OND precipitation for Miami (FL).

Alabama and Georgia in the winter (Sittel 1994). In addition, these two locations are in different climatic regimes.

Graphic diagnostics were used to determine if the sets of conditioned series were consistent with the forecast scenarios for which they were created. The distributions of quarterly precipitation totals from the synthetic series were compared to the corresponding historical distributions to check for changes with respect to climatology. A
distribution of values from a resampled set of synthetic series conditioned on climate forecasts may have a shifted central tendency or decreased variance with respect to the climatological distribution (Potgieter et al. 2003). Shifts in distributions of total precipitation from conditioned synthetic series generated using weather generators have been documented in previous works (e.g. Apipattanavis et al. 2007, Wilks 2002).

The characteristics of the daily precipitation values were also assessed. Previous works have shown that conditioning of synthetic series upon climate information (such as seasonal forecasts) with the use of weather generator procedures leads to changes in properties of precipitation such as wet day probabilities (e.g. Briggs and Wilks 1996; Wilks 2002). It was hypothesized that this resampling approach (if using non-climatological forecasts) would create sets of synthetic series which had different statistical properties than the historical data.

Since the forecasts covered seasons, the diagnostics were done for the same three-month periods (as compared to individual months for the assessment of the generator). The resampling procedure results are discussed for resampling based on formats I and II separately. In addition, there were five different forecast scenarios initially utilized in the testing (see Table 2.2). The forecasts with the largest deviations from climatology were those in the very wet and very dry scenarios. In the diagnostics covered in this chapter, only the results for the very wet and very dry scenarios (and climatology) were discussed. The tercile probabilities that were used in format I and for generating the PDFs for format II for the very wet and very dry scenarios are provided in Table 3.3.
Table 3.3: Tercile probabilities for the very wet and very dry scenarios.

<table>
<thead>
<tr>
<th>Tercile</th>
<th>Very Wet Forecast</th>
<th>Very Dry Forecast</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>JFM</td>
<td>AMJ</td>
</tr>
<tr>
<td>Upper</td>
<td>60%</td>
<td>40%</td>
</tr>
<tr>
<td>Middle</td>
<td>20%</td>
<td>35%</td>
</tr>
<tr>
<td>Lower</td>
<td>20%</td>
<td>25%</td>
</tr>
</tbody>
</table>

3.2.2.1 Median Quarterly Total Precipitation

The quarterly totals of precipitation from the resampled synthetic data were analyzed to determine if the new sets of synthetic series represented wetter or drier seasonal conditions. From each synthetic series, precipitation totals were calculated for each quarter, and the medians of those quarterly totals were then compared to the corresponding historical values. This diagnostic was performed on results from the very wet and very dry scenarios.

Diagnostics revealed that the medians of quarterly precipitation totals from the synthetic series changed as expected according to the forecast scenario. Boxplots of these median precipitation totals for Chipley (FL) are shown for synthetic series resampled based on format I in Figure 3.27 and for synthetic series resampled using format II in Figure 3.28. For very wet forecasts (left panels), the historical median totals lie beneath the majority of corresponding synthetic values represented by the boxplots, and for very dry forecasts, the historical values lie above the majority of synthetic values. For both formats, the historical value fell outside the contiguous ranges of the boxplots in JFM for both forecast scenarios and in OND with the very wet scenario. The diagnostics of the synthetic series from Miami (FL) (not shown) reveal similar deviations from climatology for both scenarios and formats.
The relative magnitudes of deviations in different quarters in these boxplots correspond with the magnitude of deviations from climatology of the respective non-climatological forecasts. The magnitudes of the deviations from climatology (33%-33%-33%) in JFM (very wet and very dry) and OND (very wet only) forecasts were greater than those in AMJ and JAS. Accordingly, the deviations of median quarterly precipitation totals were greater in JFM and OND (very wet only) than in AMJ and JAS.

Similar results were observed by Yates et al. (2003). They used a k-NN generator to create conditioned synthetic series by resampling the input data (as opposed to output
Figure 3.28: Boxplots of median quarterly total precipitation (HI) (for wet days only) for each conditioned synthetic series, calculated separately for each quarter for a) a very wet forecast and b) a very dry forecast at Chipley (FL) using format II. The boxplots indicate the dispersion of the statistics for the synthetic series. The circles joined by lines indicate the corresponding statistics for the historical series.

data in this work) for their weather generator at locations in the Midwestern United States. They used this same diagnostic for monthly total precipitation when resampling was based on alternative climate scenarios (warmer-wetter springs, etc.), and it was found that when resampling based on a warmer climate scenario, the medians of monthly precipitation totals deviated substantially from the historical values in some months.
3.2.2.2 Distributions of Quarterly Total Precipitation

Plots were made of the histograms and corresponding density estimates of quarterly precipitation totals of the historical data and resampled series to check for shifts in the distributions consistent with the forecasts. These diagnostics were done for both formats and for two quarters (JFM and JAS) with synthetic series from two scenarios (the very wet and very dry forecasts) and the historical data (representing a climatological scenario). JFM and JAS were selected to provide a contrast between the most and least non-climatological forecasts.

The distributions of quarterly precipitation values shifted in the expected direction for both methods during JFM. The distributions for series consistent with format I and the historical data for JFM at Chipley (FL) are shown in Figure 3.29. The three vertical lines in each panel indicate the 0.05 quantile (Q5), median (Q50), and 0.95 quantile (Q95) of the respective data sets. The values of these quantiles are provided in Table 3.4. During JFM, the distributions shift upwards when moving from a very dry to climatological to very wet scenario, and accordingly, the three vertical lines shifted with the distributions. In Table 3.4 among the numerical values calculated for these distributions, the largest changes were in Q50 between all three scenarios and in Q95 between climatology and a very wet scenario. Figure 3.30 is the corresponding plot for format II. The trends were similar and the shifts were in the expected directions when using this format. The Q95 increase from climatology to a very wet scenario was not as large as with format I.
Figure 3.29: Histograms of JFM quarterly total precipitation values (HI) (for wet days only) of a) all conditional series for a very dry forecast, b) the historical series, and c) all conditional series for a very wet forecast at Chipley (FL) using format I. The orange lines indicate a kernel-density fitted to each histogram. The vertical lines indicate the 0.05 quantile, median, and 0.95 quantile of the corresponding data sets.
Figure 3.30: Histograms of JFM quarterly total precipitation values (HI) (for wet days only) of a) all conditional series for a very dry forecast, b) the historical series, and c) all conditional series for a very wet forecast at Chipley (FL) using format II. The orange lines indicate a kernel-density fitted to each histogram. The vertical lines indicate the 0.05 quantile, median, and 0.95 quantile of the corresponding data sets.
<table>
<thead>
<tr>
<th>JFM</th>
<th>Format I</th>
<th>Format II</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Q5</td>
<td>Median</td>
</tr>
<tr>
<td>Very Dry</td>
<td>755</td>
<td>1290</td>
</tr>
<tr>
<td>Historical</td>
<td>815</td>
<td>1526</td>
</tr>
<tr>
<td>Very Wet</td>
<td>907</td>
<td>1791</td>
</tr>
</tbody>
</table>

Table 3.4: Q5, Q50, and Q95 for distributions of JFM quarterly precipitation for the historical data and different forecast scenarios.

<table>
<thead>
<tr>
<th>JAS</th>
<th>Format I</th>
<th>Format II</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Q5</td>
<td>Median</td>
</tr>
<tr>
<td>Very Dry</td>
<td>916</td>
<td>1551</td>
</tr>
<tr>
<td>Historical</td>
<td>961</td>
<td>1670</td>
</tr>
<tr>
<td>Very Wet</td>
<td>980</td>
<td>1700</td>
</tr>
</tbody>
</table>

Table 3.5: Q5, Q50, and Q95 for distributions of JAS quarterly precipitation for the historical data and different forecast scenarios.

The shifts in distributions for JAS, when the forecasts utilized deviate less from climatology, appeared to be smaller. In addition, there were negligible changes in some of the quantiles. The numerical values of Q5, Q50, and Q95 for JAS are shown in Table 3.5. For Chipley (FL) with format I (Figure 3.31), changes were noticeable in the distributions with the changing scenarios. The changes in Q5, Q50, and Q95 (Table 3.5) were greater when moving from climatology to a very dry scenario as compared to moving from climatology to a very wet scenario when the changes in quantiles ranged from only 7 HI (change in Q95) to 30 HI (change in Q50). For JAS when resampling using format II (Figure 3.32), the shifts of the distributions were also in the expected directions. However, changes at extreme quantiles, while negligible in size, in some cases were opposite of what is expected. The value of Q5 (Table 3.5) for the very wet scenario (944 HI) is slightly smaller than Q5 for the corresponding historical distribution (961 HI),
Figure 3.31: Histograms of JAS quarterly total precipitation values (HI) (for wet days only) of a) all conditional series for a very dry forecast, b) the historical series, and c) all conditional series for a very wet forecast at Chipley (FL) using format I. The orange lines indicate a kernel-density fitted to each histogram. The vertical lines indicate the 0.05 quantile, median, and 0.95 quantile of the corresponding data sets.
Figure 3.32: Histograms of JAS quarterly total precipitation values (HI) (for wet days only) of a) all conditional series for a very dry forecast, b) the historical series, and c) all conditional series for a very wet forecast at Chipley (FL) using format II. The orange lines indicate a kernel-density fitted to each histogram. The vertical lines indicate the 0.05 quantile, median, and 0.95 quantile of the corresponding data sets.
and Q95 for a very dry scenario (2586 HI) is slightly greater than the Q95 for the historical data (2581 HI). The changes in the Q50 (132 HI) when moving from climatology to a very wet scenario were somewhat larger when using format II than the corresponding changes when using format I. These diagnostics indicated that resampling may not lead to significant changes in the distributions of quarterly precipitation totals of synthetic data when based on forecasts which have only small deviations from climatology.

Other works which utilized weather generators for conditioning of synthetic series have documented shifts in distributions of total precipitation in synthetic series associated with predicted climate scenarios (e.g. Grondona et al. 2000; Apipattanavis et al. 2007). Grondona et al. (2000) created synthetic series with a weather generator by conditioning the parameters upon ENSO phase in Pergamino, Argentina. They analyzed density estimates of monthly precipitation totals for the month of November from synthetic series that had been created by conditioning on an ENSO cold event, an ENSO warm event, and a neutral event. The distribution of values for the cold event had a different central tendency (shifted lower) and had a smaller spread as compared to the climatological distribution. The distributions of values from the warm event and neutral event had similar central tendencies, while the warm event had a larger proportion of higher values.

3.2.2.3 Q-Q Plots of Distributions of Quarterly Total Precipitation

Quantile-Quantile (Q-Q) plots comparing distributions of quarterly precipitation totals from synthetic series and historical data were created to compare and contrast the results when using format I versus using format II and to illustrate how the resampling impacts different sections of the distributions. In Q-Q plots (see Figure 3.33), the points
represent the comparison of a range of estimates in the two distributions. If the distributions are similar, then points should fall near the 1:1 line (slope = 1). If points are above the 1:1 line, then the distribution of values represented by the y-axis (the simulated sequences) is displaced to the right of the distribution represented by the x-axis (the historical data). This indicates that the values in the distribution represented by the y-axis are relatively larger than values at corresponding points in the distribution represented by the x-axis. The reverse is true if points were below the 1:1 line. For this diagnostic, a Q-Q plot compares for only one quarter the historical data and the synthetic series from one scenario and for one format.

For the very dry scenario, the Q-Q plots revealed that in general resampling with either format I or format II shifts all sections of distributions in the expected directions. In these plots, the points should be below the 1:1 line revealing that the distributions of synthetic values were shifted to the left of the historical distributions. The Q-Q plots in Figure 3.33 are for the four quarters for a very dry forecast with format I for Chipley (FL). In these plots, most points were below the 1:1 line except at the extremes (because of the small number of extreme values in the historical data set this is ignored). There were a few noticeable places where the points were above the 1:1 lines. In the AMJ plot (upper right panel in Figure 3.33), the location of most points at values less than 1000 HI above the 1:1 line indicated that this portion of the synthetic distribution (for values less than 1000 HI) is shifted to the right of the corresponding portion of the historical distribution. The Q-Q plots when using format II for a very dry forecast are shown in Figure 3.34. Here, the location of all points (except at some extremes) below the 1:1 line
indicated that the entire distributions were shifted as expected for all quarters. There were no noticeable deviations in these plots.

A contrast between resampling based on format I versus format II is apparent when viewing Q-Q plots comparing results from a very wet scenario. If the resampling performs as expected, all points should be located above the 1:1 line (once again, the extremes were ignored). In Figure 3.35 (format I), in JFM (upper left panel) and OND (bottom right panel), the points were above the 1:1 line. However, in JAS (bottom left panel), most of the points between 1750 and 2500 HI were shifted below the 1:1 line, indicating that these portions of the synthetic distributions were shifted lower than the

Figure 3.33: Quantile-quantile plot of quarterly total precipitation (HI) (for wet days only) for a very dry forecast at Chipley (FL) using format I during a) JFM, b) AMJ, c) JAS, and d) OND. The x-axis corresponds to all of the historical totals. The y-axis corresponds to all of the conditional synthetic totals. The line has a 1:1 slope.
corresponding portions of historical distributions at these points. There were also unexpected deviations of some points in AMJ (top right panel). These results contrast with the results for using format II. Looking at Figure 3.36 for results with format II, for all four quarters, the points were consistently above the 1:1 line except at the extremes.

Overall, these diagnostics indicated that resampling with format II may have some advantages. For both scenarios (very wet and very dry) and in all quarters, the appropriate shifts occurred in all sections of the distributions when using format II. Conversely, with forecasts having smaller deviations from climatology (in AMJ and JAS), when utilizing
format I some sections of distributions did not shift according to the respective forecast scenario.

3.2.2.4 Probability of a Wet Day

Next, the properties of daily precipitation of the conditioned series were analyzed to check for changes with respect to the statistics of the historical data. Wet day probabilities were analyzed to check for changes in the occurrence of daily precipitation.

Figure 3.35: Quantile-quantile plot of quarterly total precipitation (HI) (for wet days only) for a very wet forecast at Chipley (FL) using format I during a) JFM, b) AMJ, c) JAS, and d) OND. The x-axis corresponds to all of the historical totals. The y-axis corresponds to all of the conditional synthetic totals. The line has a 1:1 slope.
The wet day probability statistics of the sets of conditioned synthetic series changed as expected when resampling based on non-climatological forecasts. Boxplots for probability of wet days for both very wet and very dry scenarios for Chipley (FL) are shown in Figures 3.37 (format I) and 3.38 (format II). In all four quarters for the very wet scenario (left panel), for both formats, the historical value lies near or within the lower quartile of the boxplots of synthetic values. In the very wet scenarios, the greatest shift from the historical value occurred in OND with both formats. In very dry forecasts (right panels) for all four quarters, the historical value lies above the corresponding
median (horizontal line within the boxes) of synthetic values. The greatest shifts occurred
in JFM, while only slight changes occurred in the other three quarters with either format.
The diagnostics for Miami (FL) (not shown) revealed similar deviations from the
historical probabilities with both formats and forecast scenarios. Overall, these
diagnostics reveal that the wet day probabilities changed as expected when conditioning
on non-climatological forecasts. In addition, larger shifts occurred in quarters when the
forecasts were more non-climatological (OND and JFM).

Figure 3.37: Boxplots of the probability of a wet day (precipitation ≥1.2 HI) for
each conditioned synthetic series, calculated separately for each quarter for a) a
very wet forecast and b) a very dry forecast at Chipley (FL) using format I. The
boxplots indicate the dispersion of the statistics for the synthetic series. The
circles joined by lines indicate the corresponding statistics for the historical series.
Briggs and Wilks (1996) showed that wet day probabilities of conditioned synthetic precipitation values shifted with changes in forecast tercile probabilities at locations in New York. Using forecasts for both precipitation and temperature, he illustrated changes in wet day probabilities as forecast tercile probabilities shifted incrementally from climatology.

Figure 3.38 Boxplots of the probability of a wet day (precipitation ≥1.2 HI) for each conditioned synthetic series, calculated separately for each quarter for a) a very wet forecast and b) a very dry forecast at Chipley (FL) using format II. The boxplots indicate the dispersion of the statistics for the synthetic series. The circles joined by lines indicate the corresponding statistics for the historical series.
3.2.2.5 Median Daily Precipitation

Changes in daily precipitation intensity characteristics were also assessed. The median daily precipitation value was calculated by quarter for all synthetic series. This diagnostic was performed on results from the very wet and very dry scenarios.

Overall, resampling did affect the properties of daily precipitation intensity. Boxplots show the dispersion of daily precipitation intensity values for each quarter and synthetic series at Chipley (FL) in Figures 3.39(format I) and 3.40 (format II). In these figures, the median values of the boxplots of synthetic precipitation intensity values were above the historical values in most quarters for very wet forecasts. The largest shifts appeared in JFM and OND for very wet forecasts with use of both formats. The median synthetic value and historical value were the same in the very wet scenario during JAS with format I. For a very dry forecast, the medians of synthetic values were beneath the historical value in all quarters and for both formats, with the largest shifts having occurred in JFM. Diagnostics for Miami (FL) for this statistic revealed similar changes when conditioning.

Katz et al. (2003) found that statistics of daily precipitation intensity changed in sets of conditioned synthetic series produced by a weather generator conditioned upon large-scale atmosphere-ocean circulation indexes (Bermuda High and New Orleans pressure indexes) at locations in the SEUS. They found that median daily precipitation intensity during the winter was altered with changes in this conditioning index.
Figure 3.39: Boxplots of median daily precipitation (for wet days only) for each conditioned synthetic series, calculated separately for each quarter for a) a very wet forecast and b) a very dry forecast at Chipley (FL) using format I. The boxplots indicate the dispersion of the statistics for the synthetic series. The circles joined by lines indicate the corresponding statistics for the historical series.
Figure 3.40: Boxplots of median daily precipitation (for wet days only) for each conditioned synthetic series, calculated separately for each quarter for a) a very wet forecast and b) a very dry forecast at Chipley (FL) using format II. The boxplots indicate the dispersion of the statistics for the synthetic series. The circles joined by lines indicate the corresponding statistics for the historical series.
Chapter 4
Discussion and Future Work

4.1 Summary of Approach and Results

In this work, tools were developed to create sets of simulated daily weather series conditioned upon seasonal climate forecasts for the Southeastern United States. First, a weather generator developed by Apipattanavis et al. (2007) was implemented and tested for use in the Southeastern United States. Then, procedures were developed to create synthetic series conditioned on two different formats of seasonal climate forecasts. The procedures involved resampling the output of the weather generator. Such conditioned series can subsequently be used as input into process models to explore impacts of climate variability on sectors such as agriculture and water resources.

The first tool used in this work was the semi-parametric weather generator developed by Apipattanavis et al. (2007) that attempts to combine the advantages of both parametric and nonparametric methods. This model consists of two major components: a parametric first-order, three-state Markov chain for generation of precipitation occurrence, and a k-NN algorithm for generating values of weather variables. The Markov model has been shown to perform well for capturing rainfall spell statistics while the nonparametric k-NN method has had success with capturing characteristics of variables such as skewness and cross-correlation. A third precipitation state (“very wet” days) was added to the Markov chain by Apipattanavis et al. (2007) to address the undersimulation of medium to large daily precipitation amounts when using a two-state model.
Using the semi-parametric weather generator, 100 synthetic series of lengths ranging between 57 and 60 years (the same as available historical) were created at 11 different locations in the Southeastern United States. The studied locations were selected to cover the geographic and climatic range of the study area, and mostly coincide with important centers of agricultural production.

Initial testing of the semi-parametric generator revealed systematic temperature biases at all locations. To address this issue, empirical corrections were applied to the synthetic temperature output of the weather generator. The addition of corrections did not negatively impact the characteristics of the synthetic data.

A broad range of graphic diagnostics was used to assess the performance of the weather generator. Most statistics of the historical weather were successfully captured. The standard deviations, skewness coefficients, and cross correlation coefficients of synthetic daily Tmax and Tmin values were similar to the corresponding statistics of the historical data. However, there was a prominent underestimation of lag-1 correlation of Tmax and Tmin on wet days, on dry days during the summer, and on days that involve transitions between different precipitation states. The Markov chain performed well, successfully reproducing precipitation occurrence statistics such as wet day probabilities and mean and maximum spell length statistics. For precipitation amounts, distributions of daily precipitation intensity and both monthly and annual precipitation totals were well reproduced. In addition, the annual patterns of standard deviation of daily precipitation and monthly total precipitation (interannual) were adequately captured.

A few graphic diagnostics focused on the assessment of replication of extremes in the historical statistics: lengths of very hot and very cold spells (spells of five days or
more), day of first freeze and last freeze, and total number of freezing days (not
mentioned in text) were all examined. The weather generator performed satisfactorily for
these diagnostics at all locations. For precipitation occurrence, the generator performed
well in simulating the longest annual dry and wet spells at all locations. Annual extreme
rainfall events also were well modeled.

The second tool utilized was a new approach to creating sets of synthetic series
conditioned on seasonal climate forecasts by using weighted resampling of the output of
the hybrid weather generator. Previous approaches to creating conditioned sets of
synthetic series involved either modification of the parameters of parametric weather
generators or resampling the weather data used as input to the generator. However, the
approach followed in this work avoids some of the problems reported for these other
alternatives; for example, daily precipitation occurrence and intensity relationships are
unaltered.

Resampling procedures were developed for two different formats of climate
forecasts currently available and produced routinely by operational agencies. These
formats are (a) the commonly used tercile format and (b) a probability distribution
function (pdf) similar to the probability of exceedance forecasts issued by the CPC. The
forecast in pdf format provides more information than the three values of the tercile
format forecasts, and thus the resampling based on the pdf format should create series
which capture more details of the forecast. Both the pdf forecast and the dense
distribution of synthetic values can easily be subset at finer intervals than three terciles
for resampling.
Diagnostics revealed that the resampling approach successfully created conditioned synthetic series consistent with the forecast scenario of interest when resampling was based on either forecast format. Distributions of quarterly total precipitation of the synthetic series were found to be shifted with respect to the historical distributions, and in general the occurrence and intensity statistics of daily precipitation changed with respect to the historical statistics. In addition, Q-Q plots confirmed that in some cases resampling based on the pdf format did capture more details of the forecasts as compared to resampling based on tercile formats. Finally, it was observed that prominent changes in the statistics of daily precipitation only occurred when resampling was based upon forecasts which had large deviations from climatology.

4.2 Significance of This Work

The use of historical weather as input for process models (e.g., crop simulation models) has a fundamental limitation: the historical data provide only one realization of the weather process. However, processes such as plant growth show a highly nonlinear sensitivity to the arrangement of daily weather (e.g., lengths of dry or wet spells). The multiple synthetic series generated by the weather generator provide multiple equally-likely weather series but with different sequences of daily values; in this way, they are useful for a thorough risk assessment exercise.

Results from this thesis suggest that the weather generator implemented has potential for applications in the SEUS. The success of this model in capturing many statistics of the historical climate of the region should give potential users confidence in its use. For agricultural applications in the SEUS, the number and timing of freezes and droughts are extremely important: the model successfully captured the historical statistics
of both freezing days and dry spells at all locations in the study region. Synthetic sequences from this weather generator coupled with process models can provide information for decision makers to determine which management techniques would minimize losses due to unfavorable weather conditions.

Stakeholders are significantly interested in assessment of extreme events, as large impacts on society often are associated with such occurrences. For example, very hot days can impact energy demand through increased use of air conditioning, human health due to increased mortality rates, agricultural production because of heat stress on crops and cattle, and water demand. In addition to the damaging effects of freezing days upon agriculture, very cold days also can impact energy demand and disrupt transportation. This generator was shown to successfully model many extremes in the SEUS including hot and cold spells.

The importance of food security across the planet has received more attention recently as food prices have increased across the planet (http://www.fao.org/worldfoodsituation). In the United States, the importance of climate variability to agriculture has been documented as the majority of crop failures in this country have been associated with the lack of or excess of rainfall (Ibarra and Hewitt 1999). In the SEUS, the importance of climate and its effect upon society was highlighted recently due to the water shortages that occurred during the summer and fall of 2007 (http://www.gaepd.org/Files_PDF/news/Level_4_Drought_news_release.pdf). These problems are likely to be exacerbated by climate change in the coming decades. The increase in demand by decision-makers for information about climate, especially the extremes, will likely increase due to these factors.
A procedure introduced in this paper enables the creation of sets of daily weather sequences which contain information about forecasts of climate variability. Since these weather sequences are readily available as input into process models, it enables climate information to be put into production and economic outcomes. Stakeholders can then readily assess potential impacts on their sector of interest based on different management decisions. Since a distribution of sequences can be provided, information about changes in the probability of different outcomes can be assessed. In addition, increased potential for extreme events in precipitation can be discerned from these distributions. Thus, increase in the likelihood of, for example, floods or droughts due to interannual climate variability can be simulated, and decision makers can plan accordingly.

Providing climate information for crop model predictions may allow stakeholders to reduce risk to food security in certain situations. Crop model predictions based on forecasts of future climate variability could be scaled back up to provide regional assessments of food production. For example, Hansen et al. (1999) noted that winter tomato yield was adversely impacted during El Niño phases. Based on a particular forecast, changes in likely yields of tomatoes and other crops could be scaled back up to a regional level, and adjustments could be made to offset potential reduced yield of tomato in the SEUS during certain ENSO conditions.

4.3 Future Work

One possible line of future work could be to focus on the systematic temperature biases detected in the synthetic series. The generation of these biases indicates that the k-NN algorithm used in this work could potentially be improved to more accurately reproduce series of Tmax and Tmin in the SEUS. However, it should not be assumed that
this model has performed poorly compared to other works. An extensive array of
diagnostics was used in this work, and the number and breadth of diagnostics were more
extensive than what is typically done in publications regarding weather generators. For
example, many works do not assess Tmax and Tmin biases separately by month and state
(dry and wet). Thus, it is possible that the magnitude of temperature biases observed in
this work is not atypical. In addition, when like diagnostics about temperature generation
can be compared (skewness, means, etc.), on average, this model appears to have
performed similarly to other recent works for temperature generation.

While no conclusive theory was found for the cause of the systematic temperature
biases, detailed analysis of the synthetic data revealed that the k-NN algorithm sometimes
did not reproduce the historical characteristics of temperature on transitioning days and
on the dry days immediately following transitions. It is hypothesized that the inability of
the model to capture the characteristics of temperature in the historical data in these
situations is the cause or a partial cause of the temperature biases. The underestimation of
lag-1 correlation for transitions is consistent with this hypothesis. In addition, this
hypothesis could explain why systematic biases are opposite on dry and wet days.

Potential modifications to the k-NN algorithm to address this issue would involve
using precipitation occurrence characteristics as additional criteria for determining
potential neighbors. These types of modification are possible in the semi-parametric
generator since the series of precipitation occurrence for each synthetic year is
determined before the nearest neighbor process begins. For example, in addition to
limiting potential neighbors based on the precipitation state of a given day in the
historical data and the state of its successor, the precipitation state for the day following
the successor (what could be considered the second successor) could be used as an additional criterion (for example, instead of finding wet days followed by dry days, the model would search for wet days followed by dry days followed by wet days). Alternatively, the precipitation state for the day preceding the precursor could be a criterion. These types of modifications would allow the model to not only simulate days in transitions separately, but simulate the days preceding and following transitions separately. Hypothetically, this could improve simulation of transitioning days and those days following the transitions where the historical characteristics of temperature are not captured by the model. A drawback to increasing criteria for the k-NN process is that the number of available neighbors is reduced, and in the cases of very wet days, there may be few or no potential neighbors within the currently used window size of 15 days.
References


