

Air Force Institute of Technology

AFIT Scholar

---

Theses and Dissertations

Student Graduate Works

---

3-2003

## Data Mining Atmospheric/Oceanic Parameters in the Design of a Long-Range Nephelometric Forecast Tool

Richard F. Benz

Follow this and additional works at: <https://scholar.afit.edu/etd>



Part of the [Atmospheric Sciences Commons](#)

---

### Recommended Citation

Benz, Richard F., "Data Mining Atmospheric/Oceanic Parameters in the Design of a Long-Range Nephelometric Forecast Tool" (2003). *Theses and Dissertations*. 4288.  
<https://scholar.afit.edu/etd/4288>

This Thesis is brought to you for free and open access by the Student Graduate Works at AFIT Scholar. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of AFIT Scholar. For more information, please contact [richard.mansfield@afit.edu](mailto:richard.mansfield@afit.edu).



**DATA MINING ATMOSPHERIC/OCEANIC PARAMETERS IN THE DESIGN  
OF A LONG-RANGE NEPHELOMETRIC FORECAST TOOL**

THESIS

Richard F. Benz, Major, USAF

AFIT/GM/ENP/03-02

**DEPARTMENT OF THE AIR FORCE  
AIR UNIVERSITY**

**AIR FORCE INSTITUTE OF TECHNOLOGY**

---

**Wright-Patterson Air Force Base, Ohio**

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the United States Government.

AFIT/GM/ENP/03-02

DATA MINING ATMOSPHERIC/OCEANIC PARAMETERS IN THE  
DESIGN OF A LONG-RANGE NEPHELOMETRIC FORECAST TOOL

THESIS

Presented to the Faculty  
Department of Engineering Physics  
Graduate School of Engineering and Management  
Air Force Institute of Technology  
Air University  
Air Education and Training Command  
In Partial Fulfillment for the Requirements for the  
Degree of Master of Science in Meteorology

Richard F. Benz, BS  
Major, USAF

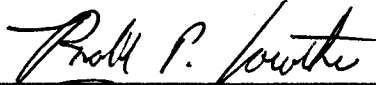
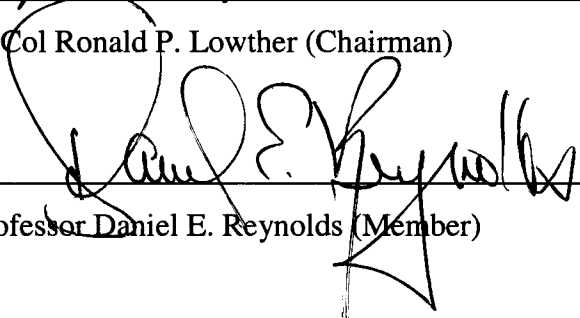
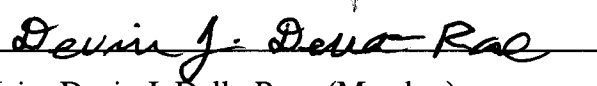
March 2003

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

DATA MINING ATMOSPHERIC/OCEANIC PARAMETERS IN THE  
DESIGN OF A LONG-RANGE NEPHELOMETRIC FORECAST TOOL

Richard F. Benz, BS  
Major, USAF

Approved:

 _____	<u>12 Feb 03</u>
Lt Col Ronald P. Lowther (Chairman)	date
 _____	<u>12 Feb 03</u>
Professor Daniel E. Reynolds (Member)	date
 _____	<u>12 Feb 03</u>
Major Devin J. Della-Rose (Member)	date

## Acknowledgments

I greatly appreciate the guidance and advice granted to me throughout my endeavors from my faculty advisor, Lt Col Ron Lowther. Professor Dan Reynolds' statistical wisdom was greatly valued. The assistance of Major Devin Della-Rose also deserves great praise. Thank you to all of the faculty, staff; and students; in particular, Captains Deleone Narcisse, Mark Allen, Rick Gonzalez, and Marc Gasbarro at the Institute for helping me achieve success. I am also grateful for the assistance provided to me by the Air Force Combat Climatology Center, specifically Mr. Dan Kerupetski, Ms. Alicia Hughes, Capt Thomas Renwick, and MSgt John Johnson, as well as, their entire support staffs.

This work would not have been possible without the support of my beautiful wife, my entire loving family and friends. Thank you everyone for your patience and understanding.

Richard F. Benz

## Table of Contents

	Page
Acknowledgments.....	iv
List of Figures.....	vii
List of Tables.....	ix
Abstract.....	x
I. Introduction.....	1
1.1 Problem Statement.....	2
1.2 Scope of Research.....	4
1.3 Research Objectives.....	4
II. Background and Literature Review.....	7
2.1 General Climate and Weather Pattern Review.....	7
2.1.1 Air Masses.....	10
2.1.2 Cyclones and Fronts.....	10
2.2 Predictors.....	12
2.2.1 Sea Surface Temperatures.....	12
2.2.2 Teleconnection Indices.....	13
2.3 Predictands.....	19
2.3.1 Cloud Parameter.....	19
2.3.2 Outgoing Longwave Radiation.....	21
III. Data Collection and Review.....	23
3.1 Collection.....	23
3.1.1 Teleconnection Index Data.....	23
3.1.2 Sea Surface Temperature Data.....	25
3.1.3 OLR Data.....	27
3.1.4 Cloud Parameter.....	29
3.2 Data Limitations.....	30
IV. Standard Statistical Methods and Results.....	32
4.1 Data Examination.....	32

4.1.1 Surface Observational and RTNEPH Data .....	32
4.1.2 OLR Data .....	34
4.2 Standard Statistical Analysis.....	38
4.2.1 Simple Linear Regression .....	38
4.2.2 Multiple Linear Regression.....	39
4.3 Removal of Seasonal Affects.....	39
4.4 Monthly Regression Results .....	40
4.5 Overall Result of Standard Statistical Analysis .....	41
V. CART Overview, Methods, and Results.....	42
5.1 CART Overview .....	42
5.1.1 Tree Splitting Rules .....	44
5.1.1.1 Gini Impurity Function .....	44
5.1.1.2 Twoing Function.....	44
5.1.2 Priors .....	45
5.1.3 Improvement Score.....	46
5.1.4 Pruning.....	46
5.1.5 Cross-validation .....	47
5.2 CART Methodology and Results.....	47
5.2.1 Classification Trees.....	49
VI. Conclusions and Recommendations .....	67
6.1 Conclusions.....	67
6.2 Recommendations.....	69
Appendix A. Correlogram .....	70
Appendix B. Linear regression results.....	71
Appendix C. CART results for Baghdad, Iraq.....	82
Bibliography .....	85
Vita.....	88



## List of Figures

Figure	Page
1. A flow-chart diagram of the research process.....	6
2. Summer 850mb geopotential heights.....	8
3. Fall 850mb geopotential heights.....	9
4. Winter 850mb geopotential heights.....	9
5. Fall cyclone storm tracks.....	11
6. Winter cyclone storm tracks.....	11
7. Spring cyclone storm tracks.....	12
8. Phases of the NAO pattern.....	14
9. Positive phase of the ASU pattern.....	15
10. Four regions of SSTs examined.....	26
11. Three regions of Afghanistan used to represent OLR origins.....	28
12. Map of observational sites (lettered/numbered solid circles) and RTNEPH grid boxes (light and dark stippling).....	30
13. Matched pairs test of RTNEPH means and surface observational means, which failed to capture the zero line.....	33
14. Time-line of OLR monthly values from 1950-2001 using the middle month of each season.....	35
15. Graphical display of the <i>t</i> -test showing the OLR difference between the two time period means for each of the selected months.....	36
16. Global CO <sub>2</sub> concentrations.....	37
17. An example of a splitting question imposed upon the root node in an attempt to isolate the largest class (normal).....	43
18a. An overgrown November tree.....	51
18b. November tree (continued) from Node 5.....	52

19a. November tree after first pruning.....	54
19b. First pruning of November tree (continued) from Node 4.....	55
20a. November tree after second pruning.....	56
20b. November tree (continued) after second pruning.....	57
21. Final November tree after third pruning.....	59
22. Pruned January classification tree with a misclassification rate of 48%.....	64
23. Pruned February tree with a misclassification rate of 40%.....	65
24. Pruned March tree with a misclassification rate of 44%.....	66
A1. The above correlogram shows the correlation between the monthly Afghanistan OLR observations at different time distances apart.....	70
C1. This January classification tree for Baghdad, Iraq OLR had a misclassification rate of 36%.....	82
C2. The February classification tree for Baghdad, Iraq OLR had a misclassification rate of 45%.....	83
C3. The March classification tree for Baghdad, Iraq OLR had a misclassification rate of 40%.....	84

## List of Tables

Table	Page
1. Months when specific teleconnection patterns are prominent.....	24
2. Coordinates representing the OLR regions in Afghanistan.....	28
3. Correlation values showing the strength of the relationships of predictands on an annual and monthly basis for the overlapping PORs.....	34
4. Multiple linear regression table of explained variance in the model due to the affects of seasonality in the predictand.....	40
5. Cross-validation misclassification rates of pruned classification trees and improvements over climatology for the months of September through May.....	63
6. Linear regression results of TIs vs. predictands for the middle month of each season .....	71
7a. Linear regression results of SSTs vs. predictands .....	72
7b. Regression results of SSTs vs. predictands for selected months .....	73
8a. Multiple regression results for January OLR vs. December predictors.....	74
8b. Multiple regression results for January surface observational data vs. December predictors.....	75
8c. Multiple regression results for January RTNEPH data vs. December predictors..	76
9a. Multiple regression results for April OLR vs. March predictors.....	77
9b. Multiple regression results for April surface observational data vs. March predictors.....	78
9c. Multiple regression results for April RTNEPH data vs. March predictors.....	79
10. Multiple regression results for July predictands vs. June predictors .....	80
11. Multiple regression results for October predictands vs. September predictors .....	81

## Abstract

The Department of Defense calls for long-range forecasts to aid in the planning of operations. The goal of this research was to explore the feasibility of predicting, one month in advance, the total monthly cloud cover over the country of Afghanistan. In an attempt to reach this goal, the following objectives were achieved: 1) climatological synoptic study of Afghanistan; 2) survey of Real Time Nephanalysis, outgoing longwave radiation (OLR), and surface observational data; 3) examination of teleconnection indices and sea surface temperatures; 4) standard statistical analysis for prediction; and 5) classification tree analysis (CART). In addition, due to current world events, CART analysis was also applied over the country of Iraq (see Appendix C).

Data were examined using standard statistical regression techniques, including linear and multiple linear regression, and then CART analysis was used for exploring possible concealed predictive structures. Standard statistics showed a strong negative correlation between monthly average OLR and surface observational total cloud cover from the fall through spring months. However, linear regression revealed very weak relationships between the predictor and predictand variables. As well, CART results contained misclassification rates that exceeded established thresholds for operational use. Further studies using CART for atmospheric science applications should be pursued.

# DATA MINING ATMOSPHERIC/OCEANIC PARAMETERS IN THE DESIGN OF A LONG-RANGE NEPHELOMETRIC FORECAST TOOL

## I. Introduction

Long-range forecasting is a daunting task for meteorologists. Many agencies rely on extended-range climate forecasts to better anticipate the needs and behaviors of industries. As well, the Department of Defense (DoD) requires long-range forecasts to aid in the monitoring and anticipation of weather impacts upon operations, planning, and their influence on the stability and welfare of nations. For example, previous theses at AFIT which were requested by DoD agencies examined long-range forecasting with the use of sea surface temperatures (SST) and teleconnection indices (TI) to predict events like sustained winds, seasonal weather severity, and heating/cooling degree-days (Freestrom, Schroeder, Randall, 2002). Other published studies outside of DoD, included forecasted rainfall amounts in Africa (Jury, 1995) and snow squalls and ice cover over Canada (Assel & Burrows, 1992), to name just a few.

Commanders preparing for operations around the world are well schooled in the fact that victory depends upon knowing the changing weather conditions throughout the entire battle campaign. As a result, they often charge their indigenous weather units or agencies such as the Air Force Combat Climatology Center (AFCCC) to anticipate the atmospheric conditions in the area of operations well into the future. Unfortunately, there are few tools in the Air Force Weather inventory that meet the basic requirements of such operations. Therefore, the goal of this research is to develop a predictive model, by statistical and classification tree analysis (CART), for CENTCOM use in forecasting mean monthly total cloud cover over Afghanistan.

## 1.1 Statement of the Problem

Extended-range forecasts are based upon the science of macrometeorology. However, Franz Baur (1951) suggested there were two fundamental problems with macrometeorology. The first problem was whether a real *Grosswetter* (large-scale flow) actually existed. If so, it was necessary to statistically show from a lengthy period of record, whether or not the likelihood of occurrence of the meteorological phenomena that determine weather are constant (aside from the annual influence) or subject to variation (Baur, 1951). Secondly, Baur believed those statistical probabilities of the meteorological elements were proven not to be constant from year to year. Thus, some governing parameters existed which created the variations of the probabilities. Baur believed terrestrial processes (ocean currents, ice conditions, volcanic eruptions) were linked to the general atmospheric circulation but offered only causal significance or that these processes merely shaped the macrometeorological event. This research attempts to use large-scale atmospheric indicators and examine them with modern statistical techniques to produce extended-range forecasts.

SST anomalies are closely monitored and modeled by climatologists as they are known to impact global circulations and their accompanying distribution of cloud patterns and hydrometeors. Trenberth (1981) noted that changes in tropical SST patterns, as a source of atmospheric circulation anomalies, were associated largely with the location of convergence zones, rainfall, and large-scale latent heat release. Although the precise relationships of these couplings are rather intricate, it is widely known that areas of convergence exist over the warmest ocean waters. For example, evidence of large SST anomalies in the tropical and North Pacific were linked to U.S. drought conditions during

1988 and flooding events during the summer of 1993 (Ting, 1997). As well, winter studies focused on the El Niño-La Niña anomaly events, concluded that SSTs greatly influence atmospheric circulation patterns. The winter study result was proven by the use of atmospheric global circulation models with fixed tropical Pacific SST values (Ting, 1997).

Another atmospheric signal employed by scientists are global teleconnection patterns. They are relations between large-scale global pressure centers, acknowledged through the presence of geophysical methods, by statistical correlations (Glantz, 1991). The origin and distribution of these pressure patterns are forced by seasonal changes of insolation linked with responses of ocean bodies and landmasses to heating. The response of oceans to sea surface temperatures, driven by the sun's energy is small due to the oceans' greater ability to retain heat. Unlike oceans, landmasses have a lesser capacity for storing heat. Thus, land masses tend to remain colder in winter and warm up noticeably in summer months. This results in pressure gradients associated with the dynamic response to the land-ocean temperature and heating disparities. These pressure gradients strengthen circulation patterns thus illustrating the symbiotic relation between the atmosphere and the oceans (Trenberth, 1981). A popular teleconnection is that of the Southern Oscillation (SO). The SO refers to an oscillation in the difference of sea surface pressures between Darwin, Australia and the south-central Pacific at Tahiti. When the waters of the eastern Pacific are abnormally warm (an El Niño event), sea level pressure drops in the eastern Pacific and rises in the western Pacific near Darwin. The reversal in the pressure gradient is accompanied by a weakening of the low-latitude easterly trade winds. This index has been related simultaneously to severe droughts and floods across

several continents (Ting, 1997). Although these pressure oscillations largely influence the weather of the tropical regions, their affect elsewhere is less understood due to the seasonal changes and the annual variability of the SO.

### *1.2 Scope of Research*

In order to provide DoD with a predictive model for cloud cover over Afghanistan, this research is structured to first understand the climatic controls and general circulation of the air masses that affect Afghanistan. Knowledge of the overall weather patterns and the annual movement of weather systems through the region is vital to understanding cloud cover patterns that exist over the area. After all possible methods of standard statistical analysis are exhausted, a data mining technique using CART, a form of binary recursive partitioning, is used to further determine if any predictability exists with present knowledge.

### *1.3 Research Objectives*

This research examines the fluctuation of SSTs and atmospheric circulation patterns and their effect on the mean cloud cover over Afghanistan. Employing the use of standard statistical methods first, and then classification trees, this study creates a predictive model for forecasting one month in advance, the trend in cloud cover over Afghanistan. If the results were promising, the same methodology would be applied to the three regions of Afghanistan as well. This study examines the feasibility of using the mean monthly total cloud cover parameter derived from the Real Time Nephanalysis (RTNEPH) model and from surface observations. As well, the prospect of using



outgoing longwave radiation (OLR) values as a proxy for cloud cover is considered. These predictands are then compared to SST values and known global teleconnection indices (TIs) (predictors) to produce the predictive model. A flow chart is provided to help illustrate the research process (Figure 1). The following specific objectives that are necessary to achieve the overall goal of this research are:

1. Perform a climatological overview of the general circulation and air masses affecting Afghanistan.
2. Define the TI and SST indices and the cloud parameters pertinent to the study.
3. Collect SST data, RTNEPH model data, OLR data, surface observational data, and TI data.
4. Perform a thorough statistical examination to compare the predictands to the global TIs and SSTs.
5. Use data mining CART analysis on the SST and TI data (predictors) by mining both sets for predictive relationships of the predictands. Develop forecast decision trees to assist in choosing particular teleconnection indices and/or SSTs that are suitable predictors (in other words, the indices with the best forecast relationship to the observed total cloud cover or OLR data).
6. After detecting all statistical relationships that may exist, produce a predictive model for CENTCOM to use in forecasting total cloud cover over Afghanistan.

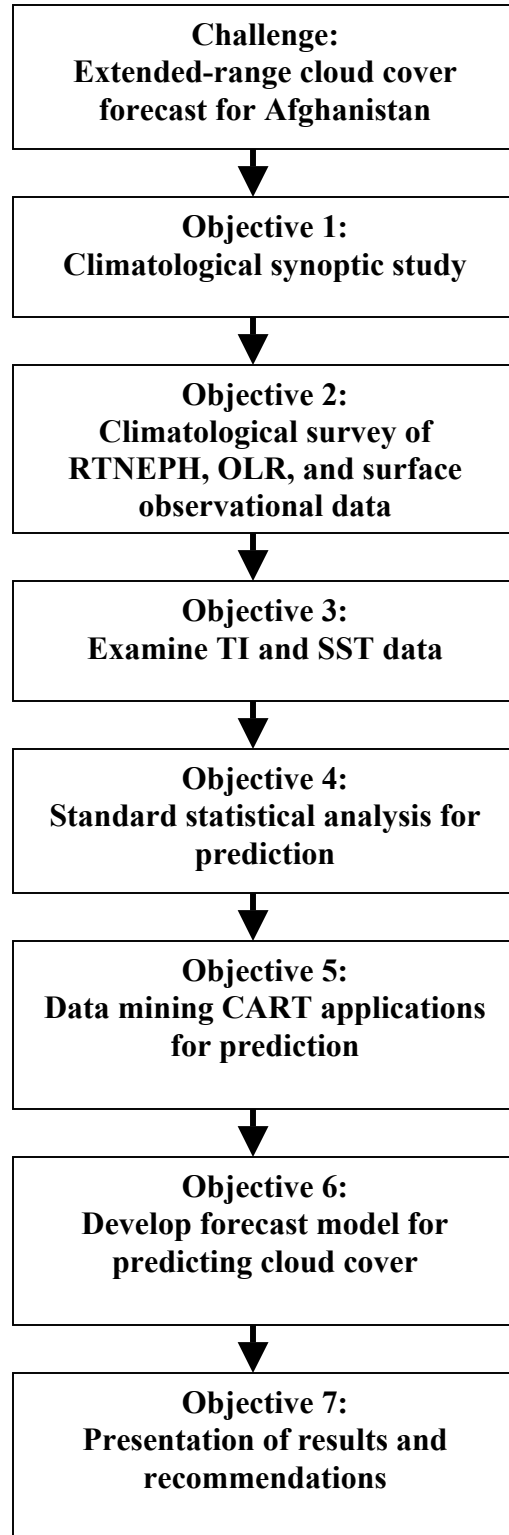


Figure 1. A flow-chart diagram of the research process.

## II. Background and Literature Review

### 2.1 General Climate and Weather Patterns Review

Knowledge of the climate and weather patterns must be gained in order to appreciate the cloud cover conditions existing over Afghanistan. An understanding of sea surface temperatures, teleconnection indices, the derivations of cloud cover values, and OLR is also required.

Afghanistan consists of mainly arid to semi-arid climates. Summers are very hot except in the high mountains as a thermal low prevails over southern Asia brought on by the high sun angle during this time of year (Figure 2). The thermal low produces hot, cloudless, dusty conditions over most of the country except in the highest mountains, over which cirrus and some cumulus clouds are found. Occasionally in the eastern parts of Afghanistan, moist, monsoonal air infiltrates from the Arabian Sea, providing low-level moisture and cloudiness.

The fall months experience a change from the average pressure patterns of summer to winter, as the thermal low decreases in intensity due to the low angle of the sun as it retreats equatorward (Figure 3). During the winter, an area of intense, semipermanent high pressure resides over the Asian continent interior (due to the vast landmass) bringing extremely cold temperatures to the region (Figure 4). The intensity and coverage of the high pressure system gradually decreases with the passage of the spring months and the onset of the thermal low that appears again over the southern portion of Asia in the summer months. In a general sense, the mean pressure decreases from north to south, often resulting in a northerly flow. Yet, during winter and spring,

the passage of low-pressure systems disrupts the general pressure distribution and flow. The frequent passage of pressure systems draws the warm dry air from the south, cold air from the north and moist air from the west (Mediterranean and Caspian Seas) (NIS, 1970).

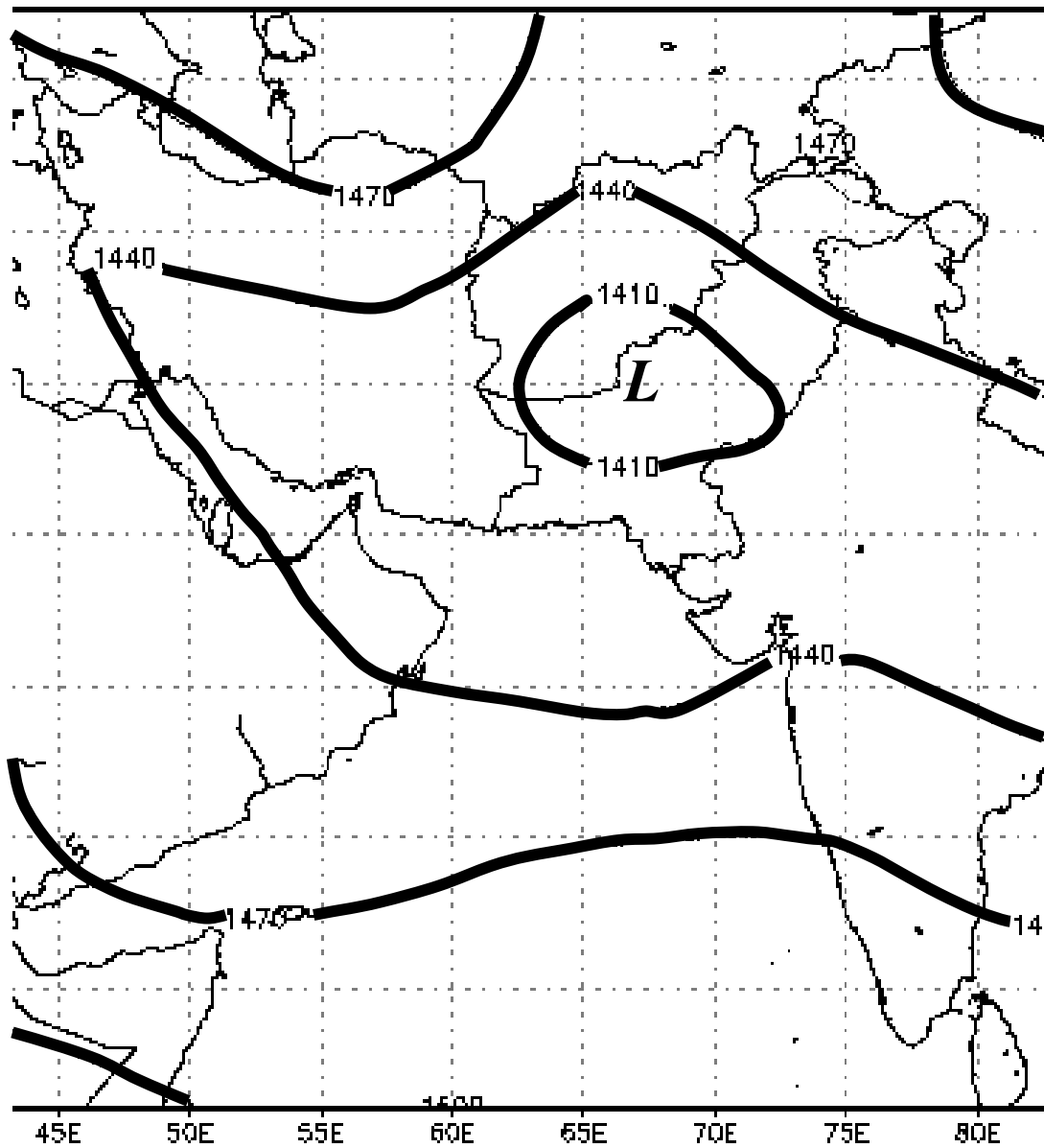
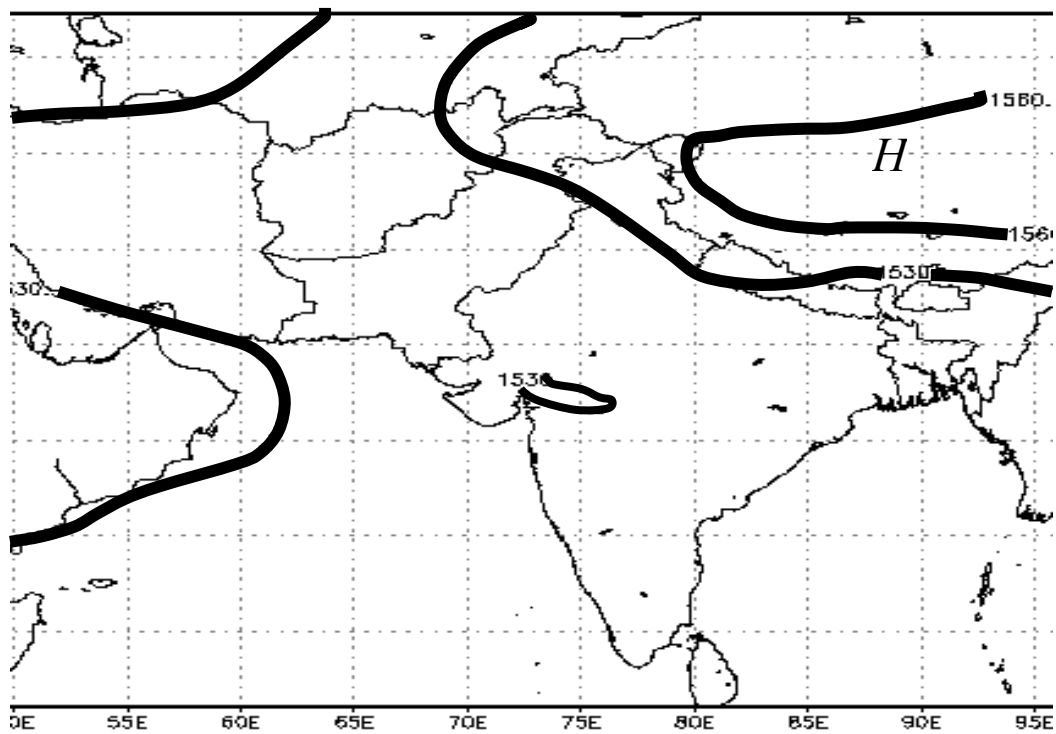
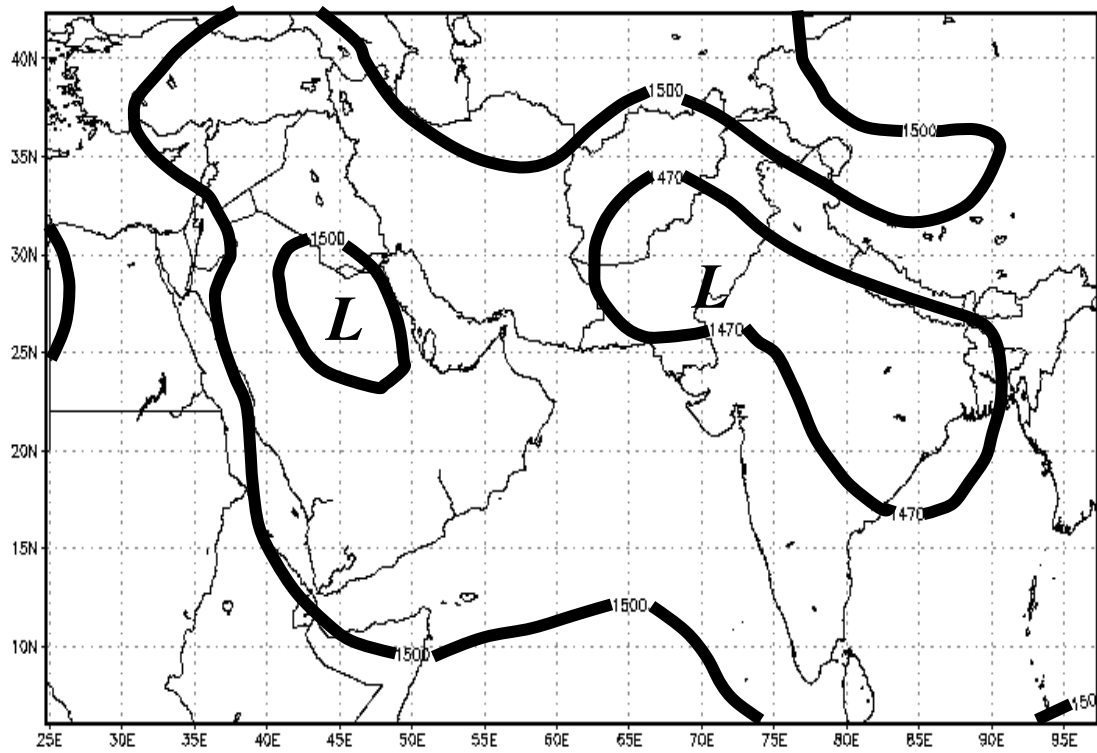


Figure 2. Summer 850mb geopotential heights (modified from FLENUMMETOC DET, 2002).



*2.1.1 Air Masses.* Air mass variability is greatest starting in late fall to early spring as low-pressure systems and accompanying fronts alter the weather patterns. Cold air enters Afghanistan from the Mediterranean region, Eastern Europe and Central Asia. The cool moist air from the Mediterranean and Eastern Europe normally follows in the wake of depressions from the west. Air originating from Eastern Europe is generally colder due to the higher latitudes with trajectories over land. Some moisture is added as the air masses cross over the Black Sea (NIS, 1970). Air masses that move southward from Central Asia are usually colder and pick up moisture over the Caspian Sea, yet most cloudiness and precipitation is depleted by the time it reaches the high mountain ranges of Afghanistan. This is due to the orographic lifting, where rising air cools and expands, forming clouds. As the air continues to rise, the water vapor continues to condense and produce precipitation. Warm, dry air masses traveling eastward from Iraq and the Arabian Peninsula also affect Afghanistan when circulations accompany depressions from the west during the cooler seasons (NIS, 1970).

*2.1.2 Cyclones and Fronts.* During the fall season, much of the precipitation for the region originates in extratropical cyclones that develop in proximity to the Mediterranean basin. These rather weak depressions move eastward, and well north of Afghanistan (Figure 5). The lows often pass north of Afghanistan and slowly migrate southward through the area as the seasons evolve into winter (Figure 6) and spring. By the end of spring, most systems have returned to their northerly tracks (Figure 7). The summer months are void of such transitory cyclones (NIS, 1970). Cold fronts that often accompany the Mediterranean lows during the colder months contribute to the unsettled weather, increasing cloudiness, precipitation, and thunderstorm occurrences. Warm



fronts associated with the transitory lows usually move northeastward through the area but are rarely associated with significant rainfall (NIS, 1970).



Figure 5. Fall cyclone storm tracks (modified from NIS, 1970).



Figure 6. Winter cyclone storm tracks (modified from NIS, 1970).



Figure 7. Spring cyclone storm tracks (modified from NIS, 1970).

## 2.2 Predictors

*2.2.1 Sea Surface Temperatures.* The oceans impact global atmospheric circulations by altering the amount of tropical/subtropical convection, which force large quantities of moisture into the upper levels of the atmosphere to be deposited elsewhere on the planet. SSTs impact the atmosphere through the release of latent heat during evaporation, and can influence the variability of the climate of a region (e.g. El Nino). For these reasons, SSTs are examined over the globe and indices are created based on these temperature anomalies. For example, in 1988 the U.S. experienced the worst drought on record resulting in tens of billions of dollars in agricultural losses,



contributing to thousands of deaths from heat stress and numerous forest fires. After a thorough analysis of the drought and how it progressed, it was suggested that the primary cause was the alteration of the atmospheric circulation across North America brought about by changes in SSTs in the tropical Pacific (Trenberth, 1991).

*2.2.2 Teleconnection Indices.* There are several methods employed in calculating such indices and the SOI is computed by a simple formula of the standardized difference between the pressure in western Pacific at Darwin (D), Australia and the pressure in south-central Pacific at Tahiti (T) (U.S. CPC, 2002):

$$SOI = \frac{\frac{\text{actual T pressure} - \text{mean T pressure}}{\text{standard deviation T}} - \frac{\text{actual D pressure} - \text{mean D pressure}}{\text{standard deviation D}}}{\text{monthly standard deviation of D and T}} \quad (1)$$

Other methods of computing teleconnection indices analyze assigned atmospheric variables at specific locations on the globe and correlate those values with other values within the respective domain (Barnston and Livezey, 1986). This method is repeated until the locations with the highest amplitudes are found and labeled as the ‘centers-of-action’ of low frequency variability. Generally, several core centers-of-action are used within a pattern, with the greatest isolated correlations being negative (Barnston and Livezey, 1986). The third method is the rotated principle component analysis (RPCA), which incorporates the use of eigenvectors. Eigenvectors are non-zero vectors which have their vector space linearly modified onto a vector which is a real number product multiplied by the original vector (Merriam-Webster, 2001). Eigenvectors of the cross-correlation matrix, which come from the time differences in the grid-point values of the selected meteorological parameter, are then scaled according to the amount of total variance they explain. They are then rotated or linearly transformed with certain

constraints to derive the major circulation patterns (Barnston and Livezey, 1986).

Readers are referred to the writings of Wilks (1995) for an in-depth explanation of the mathematics behind RPCA, as it is not the focus of this research. To date, there are fourteen common indices, not including the previously mentioned SOI, employed at the Climate Prediction Center (U.S. CPC). Each index is briefly reviewed in the following.

The North Atlantic Oscillation (NAO) pattern is a year round entity (Figure 8). It consists of a north-south dipole of 700mb pressure anomalies. One is centered over Greenland and the other of opposite sign covers the central latitudes of the North Atlantic Ocean between 35-45 degrees north. Its positive phase reflects below normal heights and

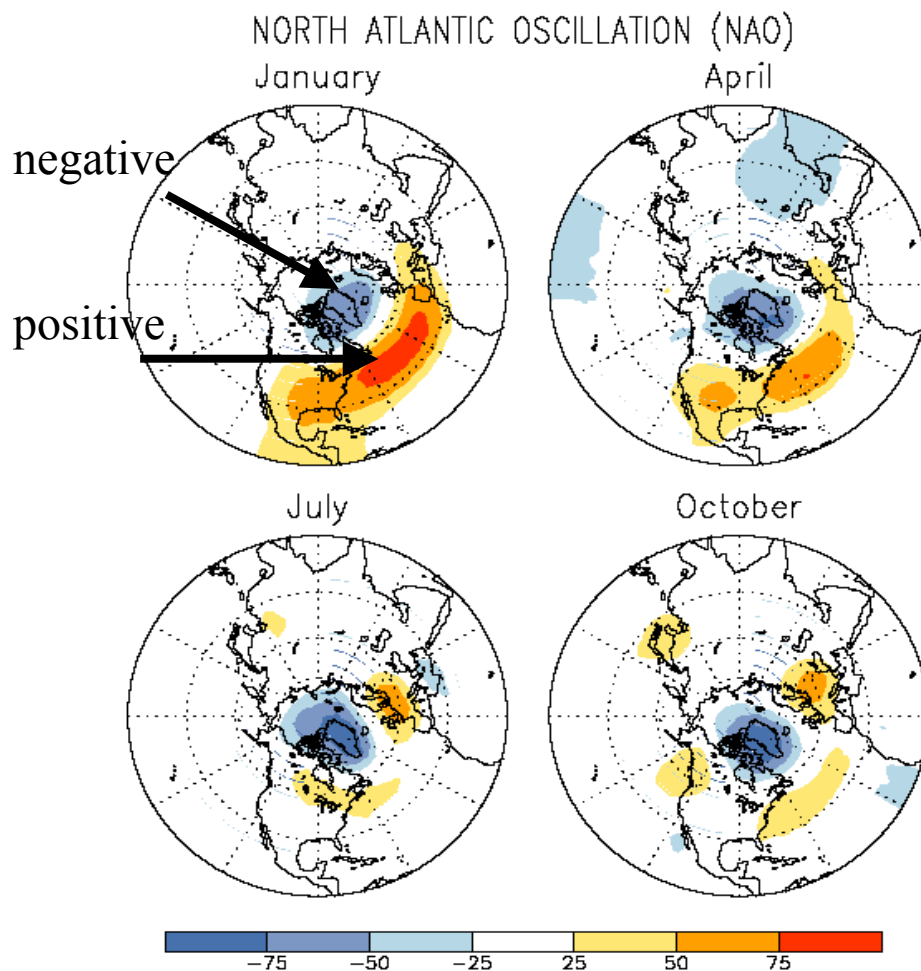


Figure 8. Phases of the NAO pattern (modified from U.S. CPC, 2002).

pressure across the high latitudes and above normal heights and pressure over western Europe stretching across the Atlantic Ocean to the eastern U.S. These phases are linked to changes in location and intensity of the jet stream over the Atlantic as well as modulations of the patterns of moisture and heat transport (U.S. CPC, 2002). The Asian Summer pattern (ASU) is strongest and only significant during the summer months (Figure 9). Unlike the other patterns mentioned, it is monopole in nature. Southern Asia and northeastern Africa are the center points with the same sign. In the positive phase, both areas experience above normal pressure heights at 700mb. The single sign phases tend to persist for several years at a time (U.S. CPC, 2002).

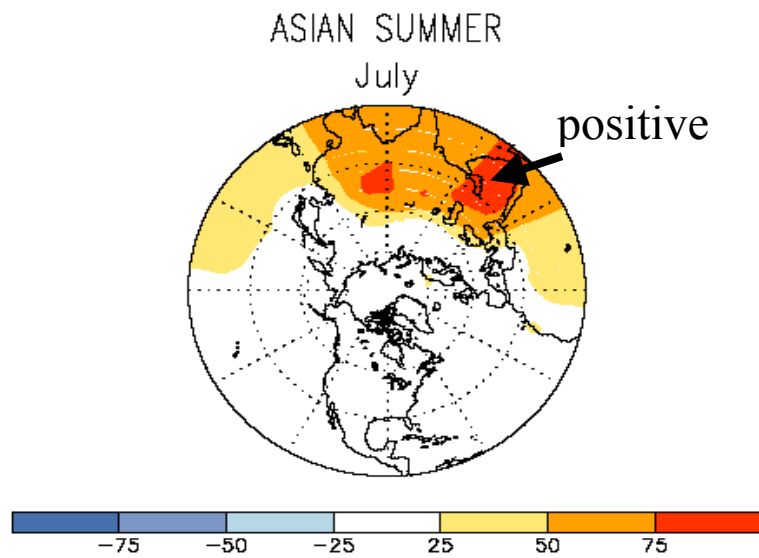


Figure 9. Positive phase of the ASU pattern (modified from U.S. CPC, 2002).

The East Atlantic (EA) is similar to the NAO structure. It appears in all months but May through August. It has a north-south dipole, which stretches across the North Atlantic Ocean. Its center in the low latitude has a strong link to the variations in the subtropical ridge's location and intensity (U.S.CPC, 2002).

The East Atlantic Jet (EAJ) pattern is the third pattern in the North Atlantic. Yet, its duration is from April through August. The north-south dipole anomaly centers are located near the eastern North Atlantic and Scandinavia, with the other near Northern Africa. During its positive phase there is a strengthening of the westerlies in the central latitudes of Europe and eastern North Atlantic. The negative phase reveals a strong split in the flow over the same regions (U.S. CPC, 2002).

The East Atlantic/West Russia (EATL/WRUS) pattern affects most of Europe and Asia during the year. Its anomaly centers are found over the Caspian Sea and western Europe during the winter. During the spring and fall, there are three centers; one over northwestern Russia, another over northwestern Europe and the third over the Portuguese coast. Negative phases appear to be more common than positive phases. The negative phase is associated with warmer and wetter conditions over Scandinavia and northwestern Russia while the opposite is seen over the Mediterranean Sea and Middle East (U.S. CPC, 2002).

The Scandinavia (SCA) pattern has a center over Scandinavia and the Arctic Ocean with two other centers of opposite sign situated over western Europe and Mongolia/western Asia. The SCA pattern exists in all but the months of June and July. The positive phase results in major blocking anticyclones over Russia and Scandinavia. The negative phase produces negative anomalies over the same regions (U.S. CPC, 2002).

The Polar/Eurasian (POL) pattern is a winter phenomenon with an anomaly center over the polar region with two opposite centers over northeastern China and Europe. It is a major indicator of the intensity of the circumpolar circulation and the associated

systems that occur over those areas. The phases tend to last for several years at a stretch. Positive phases result in below normal heights in the polar region and above normal heights over Europe and eastern Asia. The negative phase has the complete opposite affect (U.S. CPC, 2002).

A prominent pattern throughout the year is the West Pacific (WP). The north-south dipole arrangement during the winter and spring months have centers near the Bering Sea at the Kamchatka Peninsula and southeastern Asia and parts of the western North Pacific. During these months, the strongest part of either the positive or negative phase often results in variations of the location and strength of the entrance region of the Pacific jet. In the summer and fall months, the pattern becomes more wave-like and generates a third center over Alaska with an opposite sign over western North Pacific (U.S. CPC, 2002).

The East Pacific (EP) pattern is evident except in August and September. It has a north-south dipole arrangement with a northern center over Alaska and the west coast of Canada and the southern center near Hawaii. During the negative phases, a split flow is seen over the North Pacific. This often restricts the movement of the Pacific trough, confining it to the western North Pacific. The positive phase results in a deeper than normal trough in the Gulf of Alaska and positive height anomalies further south. A stronger northeastward extension of the Pacific jet stream is seen toward western North America (U.S. CPC, 2002).

The North Pacific (NP) pattern exists from March through July. One anomaly center is over the latitudes of the western and central North Pacific. The other of opposite sign is located over eastern Siberia, Alaska, and the mountainous region of

North America. During the positive phase the Pacific jet stream strengthens and moves further south from eastern Asia to the eastern North Pacific. The negative phase produces the opposite affect (U.S. CPC, 2002).

The Pacific/North American (PNA) pattern consists of four centers and exists in all months except June and July. The centers are; Aleutian Islands and southeastern U.S. and the opposite signed centers over Hawaii and the inter-mountain region of North America (during the winter and fall). During the fall and summer, the mid latitude centers appear as a wave pattern originating from the eastern North Pacific. The winter pattern shows the Aleutian center expanding over most of the northern latitudes. With the arrival of spring, the same center recedes and remains in the Gulf of Alaska (U.S. CPC, 2002).

The Tropical/Northern Hemisphere (TNH) is most active from November through February. One anomaly is center is found over the Gulf of Alaska with another of opposite sign over the Hudson Bay. A weaker center of the same sign as the one in the Alaska is found in Mexico. This pattern reveals important changes in the position and strength of the Hudson Bay Low. The TNH regulates the flow of Canadian air southward and the transport of marine air into North America (U.S. CPC, 2002).

The Subtropical Zonal pattern (SZ), as it implies, is oriented in an east-west band at 25 degrees to 35 degrees North and is the only pattern having quasi-hemispheric longitudinal extent. It is predominantly a summer pattern, often located amongst other teleconnection patterns (Barnston and Livezey, 1986).

Lastly, the Pacific Transition (PT) is strongest from May through August. A wavelike pattern of height anomalies extends from the Gulf of Alaska to the Labrador

Sea along the 40-degree latitude. A center of opposite sign can be found over the eastern U.S. (U.S. CPC, 2002).

The most relevant patterns for this research in forecasting Afghanistan cloud cover are probably the POL, SCA, EATL/WRUS, and the ASU (due to their relative proximity to Afghanistan). All teleconnection indices can be obtained from the CPC on a monthly average dating back to the 1950s.

### *2.3 Predictands*

The predictands are the forecast variables being considered for this research. They are examined in the next chapter with the most suitable variable being applied in the classification tree analysis.

*2.3.1 Cloud Parameter.* The specific parameter is simply that of the mean monthly total cloud cover. There are two sources used for obtaining this variable. The first is surface observations, which generally are recorded on a three-hourly basis at most non-U.S. observation sites. The amount of coverage is recorded as a whole number and cannot exceed 8 (as in 8/8). For aviation purposes, 1/8-4/8 describes scattered sky conditions, 5/8-7/8 is the threshold for defining a ceiling or broken sky, while 8/8 is an overcast sky. The task of obtaining adequate observations is challenging considering the remoteness of the Afghanistan area, the cultural value placed on such scientific endeavors, and extreme political instability in the region over the last 25 years or so. The surface observational data for this study was collected from AFCCC with data available from 1973-2001.

The second source of total cloud cover data comes from the RTNEPH model. The RTNEPH is a real-time cloud analysis model operated by the U.S. Air Force. It originated in the 1980's when the Air Force was looking to replace its current 3-D Nephanalysis model (established in the 70's) with something more robust. RTNEPH was designed to compile data from both satellite and conventional sources and produce an automated cloud analysis at a 25 nm horizontal resolution (Kiess and Cox, 1988). The model runs off data received from polar orbiting satellites using both visual and infrared sensors. Unfortunately, updates over certain regions sometimes do not become available until hours later due to the return time of the satellites.

The model uses a threshold technique for its analysis, in that when a satellite determines a temperature colder than its background or identifies a region brighter than the expected background, it would label that grid point with a cloud. The opposite was true for temperatures warmer than the surrounding environment and thus would label that grid point as a clear area. This method led to problems identifying low cloud types.

Once the cloud analysis was completed, the model was manually corrected for errors using satellite data over areas of specific military interest to the U.S. (Bieker, 2002). Throughout the years, RTNEPH was updated with newer algorithms and more efficient programming languages. By the early 1990's, RTNEPH was slowly replaced by another program, the Cloud Prediction Forecast System I (CDFS I). This system essentially consisted of better algorithms and the use of newer sensors onboard more orbiting satellites, to include NOAA polar orbiting satellites as well as the DMSP constellation. The POR for available data is from 1984 to the present.



In June 2002, the Air Force began using the most robust of all prior models, the CDFS II system. This version of the RTNEPH incorporates five geosynchronous and four polar orbiting satellites with multi-channel information. CDFS II also includes all available surface cloud observations to produce hourly cloud analyses. The model output currently produces a resolution of 6 nautical miles with plans at the Air Force Weather Agency (AFWA) to decrease it further. Not only does CDFS II provide cloud amount, it also determines the cloud type for each of up to four layers of clouds at each grid point. Each layer has an identified base and top measured to the nearest 30 meters up to 3000 meters, and for higher clouds, the nearest 300 meters is used. Cloud amount is given in percentages to the nearest one percent (Bieker, 2002). CDFS II provides more timely and accurate assessments of cloud data over previous versions of the RTNEPH model. In addition to total cloud cover, the model produces other variables such as cloud type and cloud amount for up to four floating layers, as well as cloud base and cloud top heights.

*2.3.2 Outgoing Longwave Radiation.* OLR is a discernable variable in the distribution of clouds and is a significant parameter in the earth energy budget and climatic variability studies (Raval and Oort, 1994). OLR is both emitted and absorbed by the atmosphere. Atmospheric gases both absorb and emit selectively at different wavelengths, roughly between 3 and 100 microns. As well, the atmosphere radiates back to space, with top-of-the-atmosphere radiation representative from the mid atmosphere (Della-Rose, 2002). Clouds behave similarly to black bodies, absorbing a portion of the longwave radiation. Thus, clouds have the ability to regionally increase the earth's albedo when compared to cloudless areas. For a specific location on the globe, the effect of a local cloud group on the regional energy balance depends on the cloud's height,

optical depth of the cloud, insolation, and the characteristics of the terrestrial surface beneath the cloud (Hartmann, 1994). The temperature of the emitting body restricts OLR, therefore polar regions and high cloud tops have lower OLR values. Higher values are found where the surface is warm with dry cloudless conditions existing overhead. Anomalies of OLR over large time periods are indicators of climatic variations caused by large-scale phenomena such as the El Niño-Southern Oscillation event (Carleton, 1991).

### III. Data Collection and Review

#### 3.1 Data Collection

Five different data sets were examined in this research. Teleconnection indices were obtained from the U.S. Climate Prediction Center. The SST data was obtained from the Fleet Numerical Meteorology and Oceanography Detachment Asheville (FLENUMMETOC), OLR originated from the NCEP/NCAR Reanalysis Project (archived at NCDC) and processed by AFCCC for this research. Surface observational and RTNEPH data sets were obtained directly from the AFCCC Database Section.

*3.1.1 Teleconnection Index Data.* The U.S. CPC is charged with calculating monthly teleconnection index values. The period of record (POR) for the indices used in this research were computed from 1950 to 2001. As previously mentioned, the SOI uses atmospheric pressure data from Darwin, Australia and the island of Tahiti (single point correlation method) to produce a standardized index; equation (1).

For the remaining TIs, the U.S. CPC uses 700mb height data to better capture the variability in the formation and amplitude of the pressure patterns associated with the annual period of extratropical atmospheric circulation (U.S. CPC, 2002). Initially, they identify which of the TIs have prominent patterns by Rotated Principle Component Analysis (RPCA), a multivariate statistical technique. Next, the amplitudes of each pattern are computed.

According to the U.S. CPC, this technique separates the primary teleconnection patterns for all months and permits time series of the amplitudes of the patterns to be

constructed. The RPCA method is considered more robust to the grid-point analysis, usually determined from one-point correlation maps. This is due to the fact that the TIs are identified based on the entire flow area, and not just from height anomalies at a few distinct locations (U.S. CPC, 2002). Table 1 is provided to show the dominant TI patterns (as defined in section 2.2.2) throughout the year, derived by the U.S. CPC using RPCA.

Table 1. Months when specific teleconnection patterns are prominent. Numerical values indicate the mode number of the pattern for that calendar month (i.e., a one indicates that the pattern appears as the leading rotated mode for the northern hemisphere during the month). If a pattern does not appear as a leading rotated mode in a given calendar month, no value is plotted (U.S. CPC, 2002).

PATTERN	DEC	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV
NAO	2	2	3	1	1	2	3	2	2	5	1	1
EA	6	6	7	6	10	---	---	---	---	8	7	5
EA-JET	---	---	---	---	6	9	7	3	7	---	---	---
WP	4	3	4	3	4	4	6	7	8	10	4	6
EP	9	10	9	10	8	3	1	1	---	---	6	9
NP	---	---	---	2	2	1	2	6	---	---	---	---
PNA	3	1	2	5	5	10	---	---	6	6	5	2
EATL/ WRUS	7	8	10	7	9	7	---	---	---	7	3	4
SCAND	5	9	8	8	3	5	---	---	10	1	2	3
POLAR- EURASIA	1	4	1	---	---	---	---	---	---	---	---	---
TNH	8	7	---	---	---	---	---	---	---	---	---	8
PT	---	---	---	---	---	8	4	4	4	---	---	---
ASIAN SUMMER	---	---	---	---	---	---	5	5	5	---	---	---

The monthly, standardized TI values are computed after the dominant mode has been determined using the RPCA method. The U.S. CPC calculates the observed amplitudes by using the least-squares regression analysis, where a best-fit line is

computed so that the squared deviations of the observed points from that line are minimized. The amplitudes are then built into a continuous time series, and standardized (mean equal to zero, variance equal to one). Patterns that do not appear as leading modes for the month are not computed (U.S. CPC, 2002).

*3.1.2 Sea Surface Temperature (SST) Data.* SST data was obtained from FLENUMMETOC and has a POR from 1954-2001. It was derived from two data sets: the Global Marine Climate Atlas (GMCA) and the Surface Marine Gridded Climatology (SMGC). The GMCA contains surface marine observations (from ships and buoys) taken from the Comprehensive Oceanographic and Atmospheric Data Set (COADS) obtained from the National Center for Atmospheric Research (NCAR) from 1954-1996. The observations are gridded into 1-degree geographical boxes stored in a simplified data format. Each observation, referenced spatially and temporally, may contain up to seventeen different physical quantities. A data file contains all observations within the 1-degree geographical box. Each observation is 46 bytes in length with each variable stored as a two-byte signed integer. In order to access the data from the files, an extraction routine in the C programming language is used. The program accepts latitude and longitude ranges directly and uses that input to determine which files are necessary for the extraction. When the bounding latitude and longitude values are specified, a geographic region is created and the extraction routine searches all the 1-degree boxes within the region and produces the desired parameter.

For this study four regions were examined based upon availability of data and/or general proximity to Afghanistan: the Mediterranean Sea (boxed by upper left coordinate to lower right coordinate), 15E by 37N and 34E by 31N; Gulf of Guinea, 8W by 4N and

7E by 11S; Indian Ocean, 50E by 15S and 65E and 30S; Arabian Sea, 60E by 25N and 75E by 10N (Figure 10). All the observed SSTs reported within the defined regions were then separated by year then month. A computed mean monthly SST in degrees Celsius was produced for each box for each year of every month in the POR.



Figure 10. Four regions of SSTs examined (modified from National Geographics, 2002).

The Surface Marine Gridded Climatology (SMGC) was used for the 1997-2001 SST POR. It uses the same 1-degree latitude/longitude global grid as the GMCA providing a resolution of about 60 km. The data source is from the Marine Atlas file archived at the National Climatic Data Center (NCDC). SMGC describes the

environment at the surface of the world's oceans for each month of the year represented (temporal resolution of one month), providing basic statistical measurements of eleven oceanic parameters. The sources for the data are from automatic observing buoys, ship reports and logs, and foreign meteorological services. The data sets are available in ASCII standard format and each record contains all the elements at a one-degree square/grid point. The media was read using the format corresponding to the record position (30-34) for the SST. All the one-degree grids in the overall latitude/longitude region are then summed and divided by the  $n$  observations to gain an overall mean SST for each month in each year. The sample size for each of the four regions range from several hundred to several thousand for each monthly mean value derived.

*3.1.3 Outgoing Longwave Radiation.* The OLR variable is from the NCEP Reanalysis data with an available POR of 1948-present. The majority of the NCEP Reanalysis data information is available online at the following links:

<http://wesley.wwb.noaa.gov/reanalysis.html>

<http://www.cdc.noaa.gov/cdc/data/ncep.reanalysis.derived.html>

AFCCC downloaded 52 years of monthly non-pressure level data and used wgrib, a gridded binary program, and its documentation at the website to extract upward longwave radiation at the top of the atmosphere ( $W/m^2$ ). Using the grid structure described at the CDC link above, the point(s) that correspond to the three areas of interest (Figure 11) were computed. Recall, if the results from the classification tree analysis were promising for the country as a whole, regional examination would be conducted. The three locations were approved by AFCCC. Due to the grid structure and the latitude/longitude of the areas, the following points were used: Area 1 (Northern Plains) 36.178N and

64.688E: midpoint average of four points (35,28), (36,28), (35,29) and (36,29). Area 2 (Central Highlands) is represented at 35.226N and 69.375E: from a single point (38,29). Area 3 (Southwestern Lowlands) is represented at 30.466N and 63.75E from an average of two points (35,31) and (35,32). Table 2 shows the latitude/longitude coordinates representing each region.

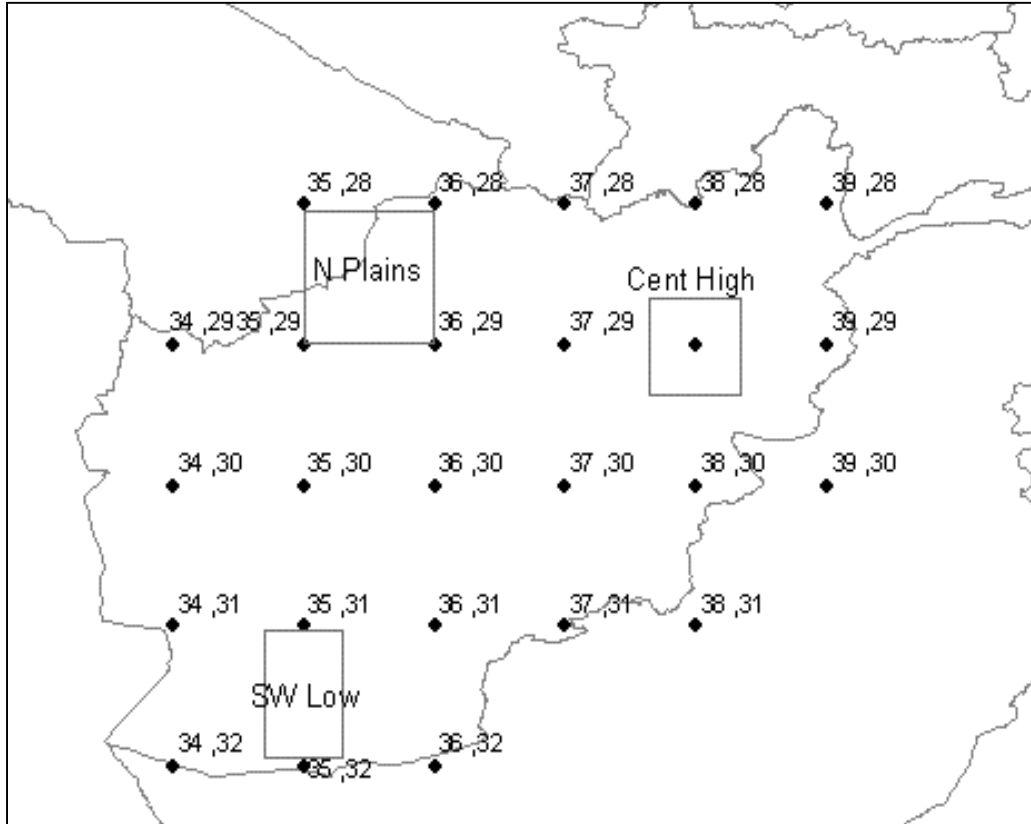


Figure 11. Three regions of Afghanistan used to represent OLR origins (modified from AFCCC, 2002).

Table 2. Coordinates representing the OLR regions in Afghanistan.

Region	I	J	Longitude	Latitude
Northern Plains	35	28	63.75	37.1305
	36	28	65.625	37.1305
	35	29	63.75	35.2264
	36	29	65.625	35.2264
Central Highlands	38	29	69.375	35.2264
Southwestern Lowland	35	31	63.75	31.4181
	35	32	63.75	29.514



After the points were identified, a Perl routine was written to use the output of wgrib to parse through the data and summarize the results. A fourth area of interest was the entire country of Afghanistan, represented by 25 points. This required a second Perl routine to be written. The OLR units are  $W/m^2$  with areas representing the country of Afghanistan and its three geographical regions.

*3.1.4 Cloud Parameter.* As mentioned previously, monthly mean total cloud cover is computed two ways, from surface observations and RTNeph model output. The surface observational data for this study was collected from AFCCC with a POR from 1973-2001. For Afghanistan as a whole, up to 27 sites were used in computing mean monthly values. For the Northern Plains, Central Highlands and Southwestern Lowlands the number of sites are 11, 12 and 4, respectively (Figure 12). The allocation of observational sites to specific regions resulted in the following:

Northern Plains: points 1, 2, 3, 4, 5, A, B, C, D, E, F

Central Highlands: points 6, 10, 11, 12, 13, 14, G, H, I, J, K

Southwestern Lowlands: points 7, 8, 9, M

The RTNeph data was derived from grid boxes 21 and 22, which are shown in Figure 12 by the dark and light stippling, respectively. There were 24 separate files (one for each month). Next, ArcView software (a powerful visualization tool which permits users to access records from existing databases and display them on maps) was used to determine the coordinates (I, J) within each box. The coordinates were then placed into the applicable regions; Northern Plains, Central Highlands, and Southwestern Lowlands. Statistical Analysis System (SAS), an analytical software code, was then created to

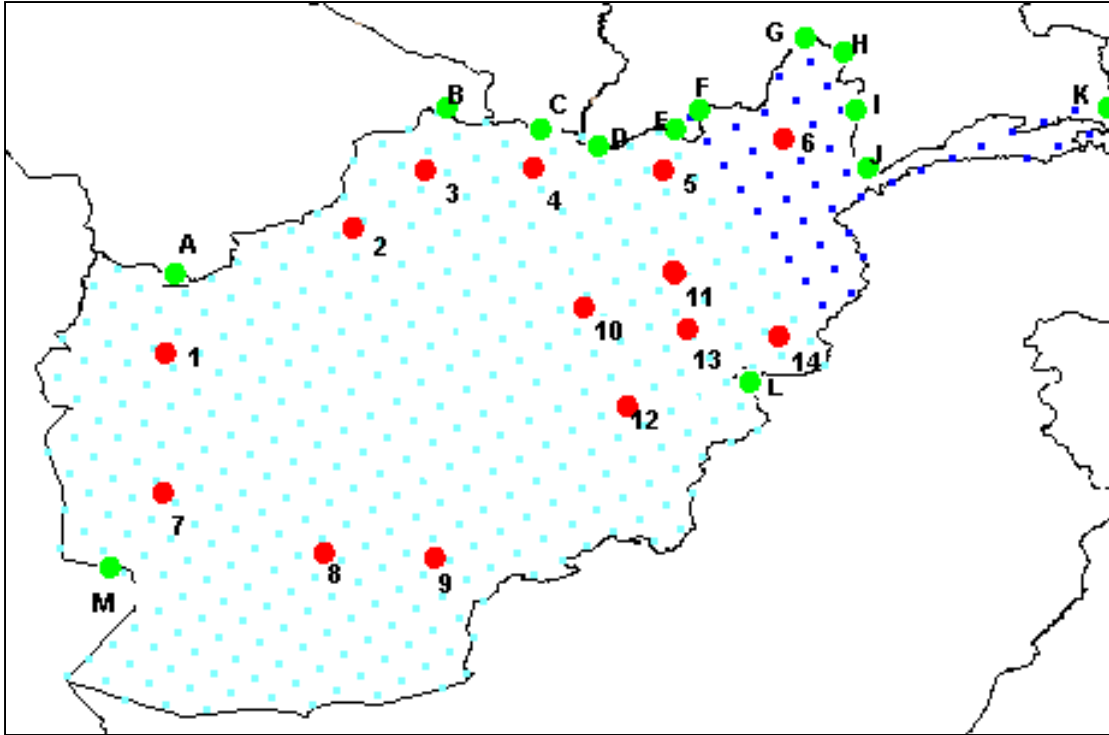


Figure 12. Map of observational sites (lettered/numbered solid circles) and RTNEPH grid boxes (light and dark stippling) (modified from AFCCC, 2002).

query the large RTNEPH files for each box. The SAS output was then transferred to an Excel spreadsheet for each of the three regions as well as Afghanistan as a whole, providing the mean total cloud cover values.

### 3.2 Data Limitations

With respect to the TIs, the 51-year POR was used in an effort to best identify trends and patterns. The exact rotation method employed by the U.S. CPC is unknown. As well, RPCA is not a popular method in analysis of planetary waves as compared to the teleconnection method, which is less detached from the original data. SST data also has a healthy POR yet issues of quality control (QC) in the oceanic measurements are unknown. The marine data files contain measurements from many sources; ship deck

logs, ship weather-reporting forms, published ship weather observations, automated marine buoys, teletype reports and foreign meteorological services. Data QC is left to computer checks, elimination of duplicate observations and manual editing of extremes.

OLR is not derived directly from any single observed variable. Rather it is derived from the model fields by the data assimilation. Thus caution must be taken when interpreting results of the reanalysis for variables of this nature (Kalnay et al, 1996).

Some of the involved variables in the radiation scheme are surface temperature, atmospheric H<sub>2</sub>O, CO<sub>2</sub>, O<sub>3</sub> and most importantly, cloud fields. The algorithm used to calculate this variable is applied across the entire POR. When new sensors or observing systems are integrated into the algorithm, a one-year, parallel reanalysis is performed to isolate perceived changes.

Total cloud cover from the surface observational data is subject to human error and consistency in the applied methodology of performing such tasks. As mentioned in the previous chapter, this data may be suspect and is hindered by the decrease in observational sites due to political instability. Also, the POR is much shorter than all but RTNeph data.

The RTNeph POR is by far the smallest of all variables. Data existed as far back as 1974 (in the 3DNeph version) but due to unknown reasons was authorized for destruction at AFCCC.

## IV. Standard Statistical Methods and Results

An examination of the data sets (using only Afghanistan as a whole and not the three regions) was necessary to aid in the explanation of the relationships between the predictor and predictand variables or reveal other issues not previously considered. Next, for predictive purposes, linear regression was applied to investigate the relationship between the two types of variables and provide a quantitative measure of their strength.

### 4.1 Data Examination

*4.1.1 Surface Observational and RTNEPH Data.* Since the surface observational and RTNEPH data sets both contained the same parameter (mean monthly total cloud cover, represented as a percentage), it was a logical assumption that the data would yield similar mean values. To compare the means of both variables, a matched pairs analysis was performed on the overlapping portion of the POR between the two data sets. This analysis uses a paired *t*-test in the examination. The test statistic, *t*, is used to emphasize that the null distribution is a *t* distribution with *n*-1 degrees of freedom rather than the standard normal distribution.

Assumptions: Null hypothesis:  $H_0: \mu_{\text{surface obs}} - \mu_{\text{RTNEPH}} = \Delta_0 = 0$

Alternate hypothesis:  $H_a: \mu_{\text{surface obs}} - \mu_{\text{RTNEPH}} \neq \Delta_0$

With a significance level of  $\alpha = 0.05$ , the *p*-value (which is the smallest level of significance at which  $H_0$  would be rejected) must be less than the alpha level of 0.05 ( $\alpha =$  probability (type I error)). Figure 13 is a plot with an x-axis equal to the mean of the two responses and the y-axis equal to the difference of the two responses. This graph is the

same as a scatterplot but rotated 45° to the right turning the original coordinates into a difference and a sum (SAS Institute Inc, 2002). If the two means were the same, the dashed lines of the 95% confidence interval would have captured the zero line on the y-axis. The confidence region failed to capture the zero line, thus revealing there was a significant difference between the two means (p-value = 0.0001). The graph also showed the RTNEPH means were consistently lower than the surface observational means. Possible reason for the RTNEPH bias may be due to the difficulty in discriminating the contrast of ice or snow-covered ground and low clouds directly above (Lowther et al., 1991). Or observers may view the sides of clouds in addition to cloud bottoms, and hence less of the sky dome is visible, as opposed to satellites which pass quickly overhead viewing the tops of clouds.

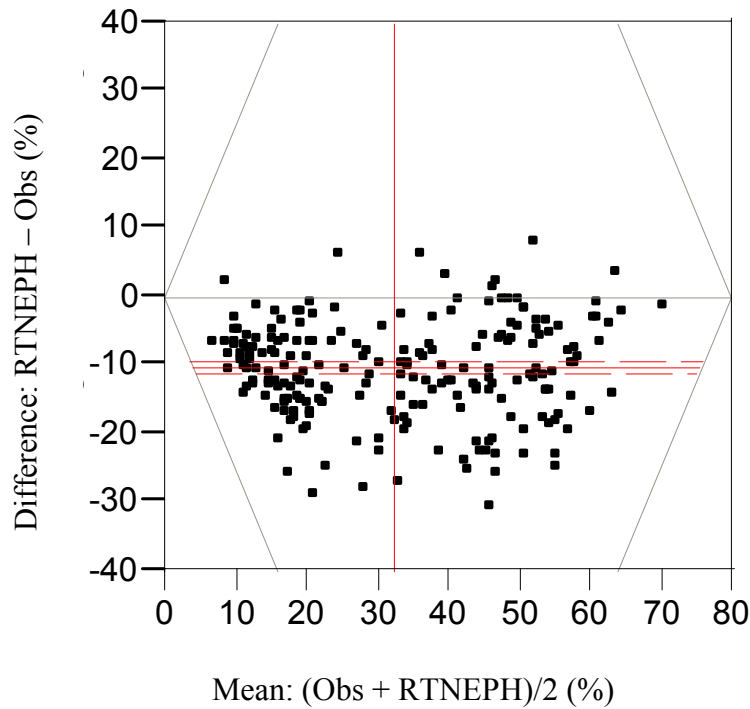


Figure 13. Matched pairs test of RTNEPH means and surface observational means, which failed to capture the zero line. Units are in percentage of total cloud cover.

4.1.2. *OLR Data.* Next, to assess the feasibility of using OLR as a surrogate or proxy for cloud cover, a correlation test (95% confidence level) was performed between the OLR, surface observational, and RTNEPH data sets, as well as, the mean of the RTNEPH and surface observational data combined. Table 3 shows the resulting r-values for an annual time frame (-0.87, -0.89, and -0.89, respectively), as well as the monthly correlation values for the overlapping PORs.

Table 3. Correlation values showing the strength of the relationships of predictands on an annual and monthly basis for the overlapping PORs.

Correlation	OLR vs Obs	OLR vs RTNEPH	OLR vs. Mean(RTNEPH and Obs)
<b>Annual</b>	-0.87	-0.89	-0.89
January	-0.69	-0.75	-0.79
February	-0.73	-0.67	-0.73
March	-0.72	-0.68	-0.76
April	-0.64	-0.74	-0.78
May	-0.80	-0.83	-0.88
June	-0.47	-0.73	-0.62
July	-0.19	-0.6	-0.48
August	-0.28	-0.82	-0.56
September	-0.23	-0.32	-0.40
October	-0.78	-0.65	-0.84
November	-0.71	-0.75	-0.81
December	-0.62	-0.85	-0.83

Overall, the results show relationships are stronger in the winter, autumn and spring months. Therefore, it may be possible to use OLR as a proxy for either variable, but the r-values for RTNEPH vs. OLR are generally stronger throughout the year. The summer

month values were less consistent perhaps due to observational error when accounting for the presence of cirrus cloud cover or cumulus cloud cover over the mountain peaks.

Afghanistan's OLR was then examined to see if there were noticeable differences over the entire reanalysis period. The middle month of each season was selected for examination. Figure 14 plots the time period in years along the x-axis and OLR in  $W/m^2$  along the y-axis. Both January and April appeared to show a distinct break in mean value. To confirm if there was a statistically significant difference, a *t*-test was performed on the two time periods (1950-75 and 1976-2001) for each selected month with a significance level of 0.05. Populations were tested for normality and equal variance.

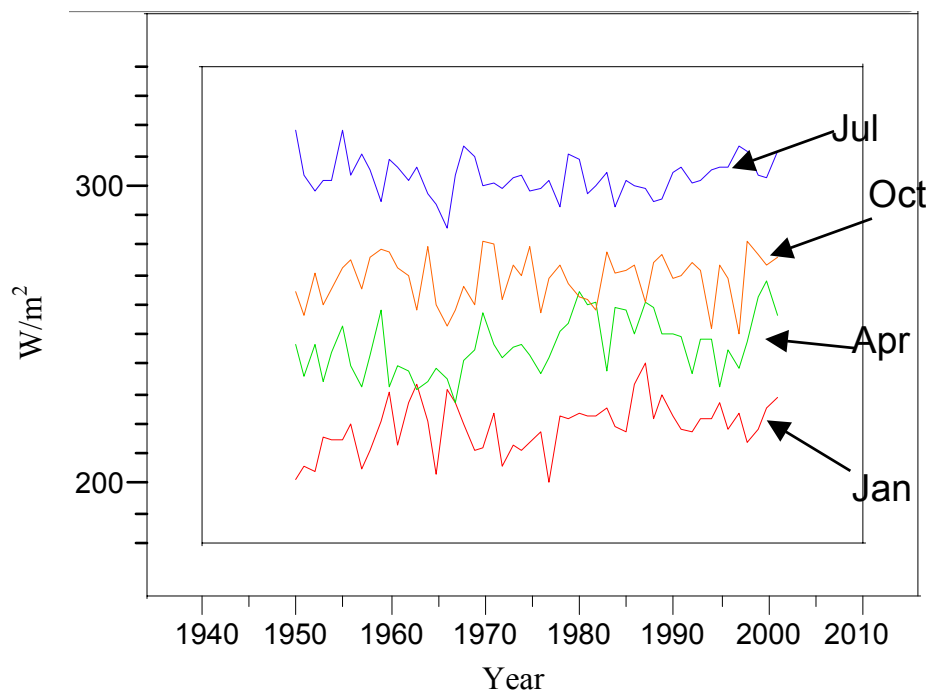


Figure 14. Time-line of OLR monthly values from 1950-2001 using the middle month of each season.

Figure 15 graphically shows the differences of the two time periods ( $W/m^2$  on the y-axis and the two time periods are levels on the categorical x-axis). The top graphs in

Figure 15 reveal the two time periods had significantly different means. The means-diamonds failed to overlap each other and the p-values were 0.0068 for January and 0.0002 for April, thereby rejecting the null hypothesis that the means are equal. The bottom graphs show the opposite result; the means were significantly similar with p-values of 0.8835 for October and 0.4133 for July. Thus, the null hypothesis of equal means was not rejected.

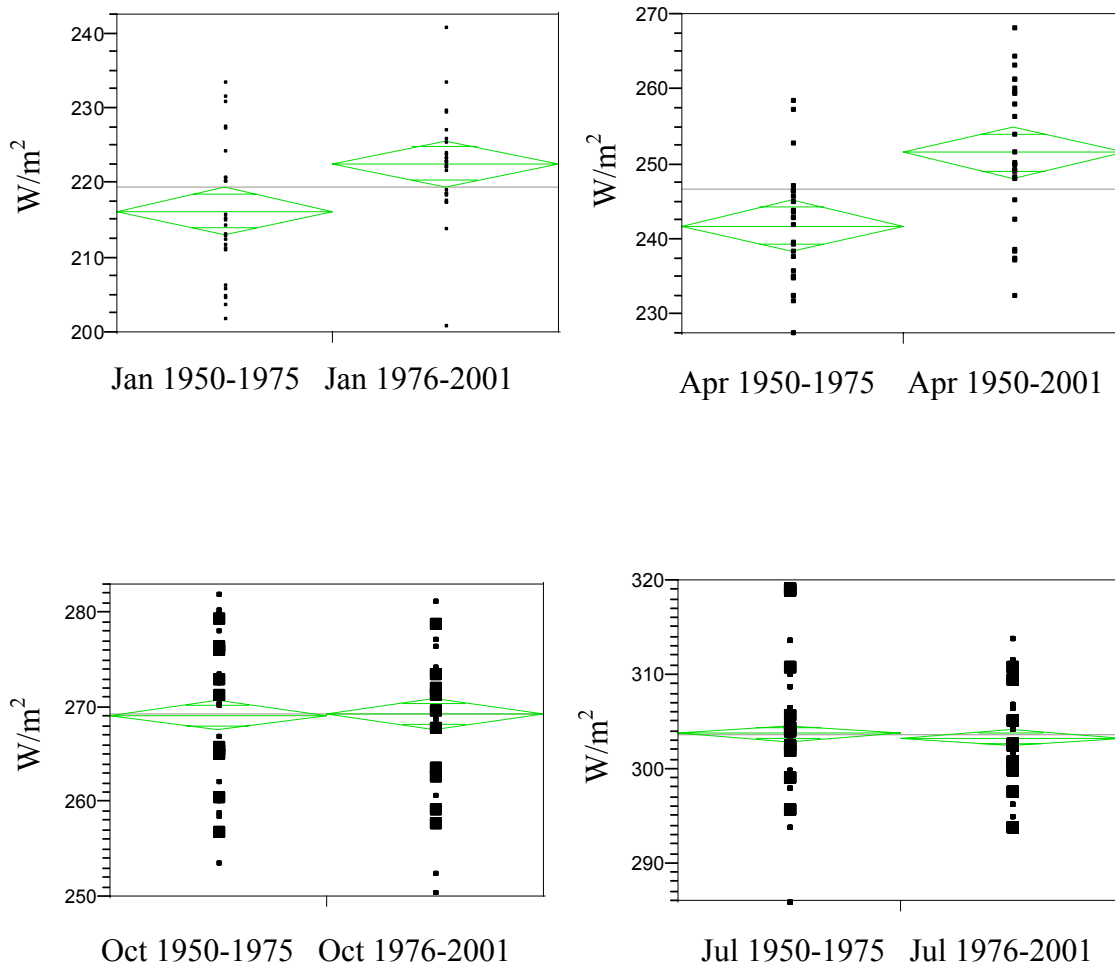


Figure 15. Graphical display of the *t*-test showing the OLR difference between the two time period means for each of the selected months. The means-diamond represent a sample mean and 95% CI. The line across each diamond represents the group mean. The vertical span of the diamond represents the 95% CI for each group. The horizontal line across the entire x-axis is the total response sample mean.



Possible reasons for the difference may be related to the global increase in CO<sub>2</sub> as seen in Figure 16. Other reasons may be due to improvements of atmospheric measurements that contribute to the OLR algorithm; improved atmospheric sounder data circa 1975; improved cloud field and wind data; overall increase in observational data sets (Jenne, 2002). However, it is unknown as to why only January and April show significant differences when all selected months are subjected to the same factors mentioned above for the same time frame.

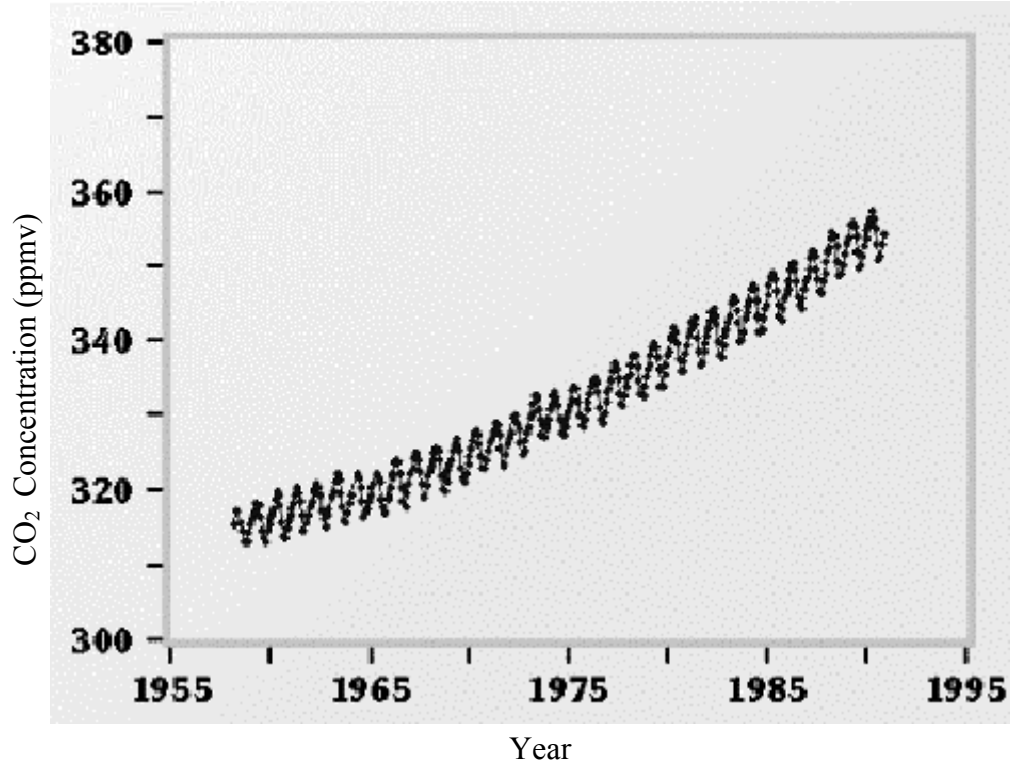


Figure 16. Global CO<sub>2</sub> concentrations. Note the slightly steeper slope from 1975-present. (modified from USGS, 2002).

## 4.2 Standard Statistical Analysis

The overall goal in using linear regression in this research is to determine if any significant relationships exist between the three predictand variables and the SSTs and TIs. Again, only the data for the country as a whole is examined and not the three regions. A brief review of simple and multiple linear regression follows.

*4.2.1 Simple Linear Regression.* The simplest way to compare an independent variable X (predictor), and a dependent variable Y (predictand) is to use simple linear regression. The goal of linear regression procedures is to fit a line through the points of a scatterplot. Specifically, computing a line so that the squared deviations of the observed points from that line are minimized. Thus, this general procedure is sometimes also referred to as least squares estimation.

A line in a two-variable space is defined by the equation  $Y = \beta_0 + \beta_1 \cdot x + \varepsilon$ . The y-variable can be expressed in terms of a constant ( $\beta_0$ ) and a slope ( $\beta_1$ ) times the x-variable plus  $\varepsilon$ , a random error term. The constant is also referred to as the *intercept*, and the slope as the *regression coefficient* or *B coefficient*. In order to establish the significance of the relationship, the value of the coefficient of determination ( $R^2$ ) is investigated, which is defined as the proportion of observed variation in the Y values that can be attributed to an approximate linear relationship between the values of Y and X or the ratio of explained variance to total variance. In other words, the  $R^2$  value is an indicator of how well the model fits the data. In order for a relationship to be determined as a viable solution to the problem, the p-value must be less than a value of 0.05 (significance level) and the  $R^2$  value should be high enough to show a strong relationship.

Baur (1951) suggested an R-value of 0.80 was necessary when using statistics for extended-range forecasting, so that the standard error of the correlation was less than 0.60 of the standard deviation. As well, Baur recommended a significance level of 0.08 or better. Thus,  $R^2$  greater than or equal to 0.64 was desired. Finally, root mean square error (RMSE) relates how accurately a model can be used to forecast results. The closer the RMSE value is to zero, the more precise a forecast is assumed to be. The individual errors are squared, added together, divided by the number of individual errors, and finally, the square root is taken. This single value summarizes the overall error of the model.

*4.2.2 Multiple Linear Regression (MLR).* The objective of MLR analysis is to construct a model relating a dependent (predictant) y-variable to more than one independent (predictor) x-variable. Thus, the model equation looks like the following:  $Y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \varepsilon$ . In order for a relationship to be determined as a viable solution to the problem, the same requirements must be met as those stated previously.

### 4.3 Removal of Seasonal Affects

As a first step in linear regression, the seasonal trends in the predictand must be acknowledged in the regression model. This issue became suspect by virtue of the results in Table 3, where the annual correlation value was larger than the monthly correlation values. Specifically, seasonality in OLR was shown by higher values during the warmer months and lower values during the colder months, which is evident in Figure 14. One method in dealing with seasonality in the predictand data is to model it by using dummy variables. Dummy variables for the monthly data have a value of either one or zero:

example, D1 = 1 if month is January, zero otherwise. To avoid multicollinearity (high correlation between the dummy variables), the number of dummy variables used is total months minus one or 11. As Makridakis et al., (1998) explained, each dummy variable acts as a new explanatory variable. The coefficients of the dummy variables reflect the average difference in the Y variable between those months and the omitted (zero value) month. For example, the coefficient of D3 (March dummy variable) is a measure of the effect of March on the forecast variable compared to February (D2). Table 4 reveals the proportion of variance in the OLR, explained by regressing on the dummy or explanatory variables is considerable and to a lesser extent, the surface observational and RTNEPH data. For example, the  $R^2$  value in the first row of Table 4 reads, the amount of variance in OLR explained by the affects of seasonality is 0.94 or 94%. A visual inspection of a correlogram (see Appendix A) may prove helpful.

Table 4. Multiple linear regression table of explained variance in the model due to the affects of seasonality in the predictand.

	$R^2$	RMSE
OLR vs seasonality dummy variables	0.94	8.21
Surface Obs vs seasonality dummy variables	0.79	8.17
RTNEPH vs seasonal dummy variables	0.77	7.94

#### 4.4 Monthly Regression Results

With the seasonality dominating the amount of explained variance, it was necessary to examine the predictor/predictand relationships on a monthly level. All three predictands were examined in the months of January, April, July, and October (Appendix B). The predictor values came from the months prior to the predictands; December,

March, June, and September. This combination was chosen due to time constraints. All four SST area values were used as well as the TIs with dominant patterns during the respective month and with general regional proximity to Afghanistan. None of the combinations showed any statistical significance, as previously set forth by Baur's recommendations.

In summary, the best results for each month follows: January RTNEPH vs. December POL pattern and Indian Ocean SST produced the highest  $R^2$  value of 0.44 with a p-value of 0.01. April RTNEPH vs. March EP pattern provided an  $R^2$  value of 0.28 with a p-value of 0.02. The June Mediterranean SST and ASU pattern vs. July RTNEPH resulted in an  $R^2$  value of 0.36 with a p-value of 0.04. Lastly, the October observations vs. the September Arab SST and the NAO pattern produced an  $R^2$  value of 0.28 with a p-value of 0.02.

#### *4.5 Overall Results of Standard Statistical Analysis*

The challenge of using SSTs and TIs to forecast a mesoscale weather parameter (Afghanistan total cloud cover) is apparent in the results obtained using standard statistics. It was shown that seasonality was the culprit in the production of high  $R^2$  values. The best  $R^2$  values when computed by month were few and not consistent in use of predictors or predictands. Having exhausted all standard statistical methods of known relevancy, data mining through CART analysis would be the next logical step of gaining any significant results for forecasting total cloud cover.

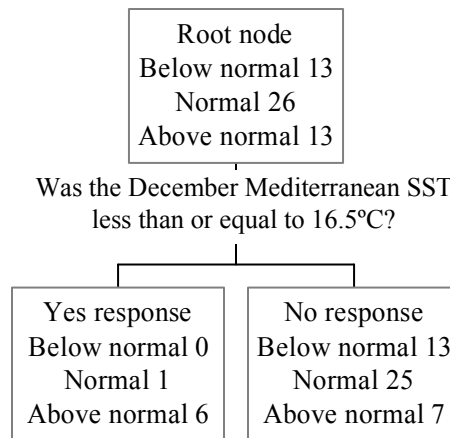
## V. CART Overview, Methods, and Results

### 5.1 CART Overview

Classification and regression tree (CART) analysis is one of the main techniques employed in data mining. Classification trees (also known as decision trees) are flow charts representing a predictive or classification model. The tree is an arrangement of simple questions whose answers outline a decision path down the tree. Due to its larger POR, OLR was chosen as the predictand. In this research, the trees are used to identify nonlinear relationships or structures between the categorical classes of the OLR variable (below normal, normal, above normal) and one or more of the predictor variables (SSTs and TIs). The objective is to build a decision tree that accurately distinguishes data among their respective categorical classes.

The foundation of this method is a binary, recursive partitioning, tree-growing algorithm developed by Breiman et al. (1984). The tree is built from the root node, which contains the entire data set of the categorical variable. The data is then split based on selected splitting rules and questions originating from the predictor variables (TIs and SSTs) to create purer subsets of the data (referred to as child nodes). In Figure 17 for example: the root node of the trees used in this study contains 52 observations of a categorical variable (such as monthly January OLR values over Afghanistan), separated into three classes; 26 observations are categorized as normal (a subjective call based on the climatology of the data set); 13 observations are categorized as below normal; and another 13 observations are categorized as above normal. A splitting rule, known in CART as the *Gini* criterion, is then imposed, in an attempt to isolate the largest class,

which in this situation is the normal category with 26 observations in it, from the other classes. Then a question is asked of the observations based on the predictor variable with the best split found, such as, “Was the December Mediterranean SST less than or equal to 16.5°C?” If the answer is yes, those observations are moved into a child node down and to the left, otherwise they are moved to the child node down and to the right of the root or original node. For instance, six above normal observations and one normal observation answer ‘yes’ to the question and are moved down from the root node to the left child node. The right node thus contains 13 below normal observations, 25 normal observations and 7 above normal observations. The left child node is now more pure than either the root node or the right child node because it contains mostly above normal observations. The right child node would most likely be split further to better isolate the observations it contains. This type of yes/no questioning recurs until the data cannot be split further.



Fi  
an  
no  
Containing the no responses.

The major problems encountered when building a tree are deciding on the appropriate splitting rules, when to declare a node as a terminal node or to continue the splitting, and validating the accuracy of the computed tree.

*5.1.1 Tree Splitting Rules.* The CART algorithm considers all possible splits for all variables in its analysis. For example: an OLR data set with 52 observations and 10 possible predictor variables would consider up to 520 possible splits. The best split is the one that produces the largest decrease in variety amongst the classes in the node (or one that produces the greatest homogeneity in the node). To assist CART in finding these best splits, there are several ‘impurity functions’ to select from to incorporate into the decision process.

*5.1.1.1 Gini Impurity Function.* The *Gini* function seeks to isolate the largest class (normal class with 26 observations, as previously mentioned) in the dataset from the remaining classes. This split searches for the best separation that produces a high amount of purity, (homogeneity or lack of variety) in the node. With *Gini*, the impurity of the node is calculated by subtracting the sum of squared probabilities of each class (below normal/normal/above normal) within the given node summed over all levels of the categorical variable. For instance, an impurity calculation for an OLR variable with a distribution of 13 below normal observations, 26 normal observations, and 13 above normal observations is as follows:

$$1-[(13/52)^2+(26/52)^2+(13/52)^2] = 0.626 \quad (2)$$

*5.1.1.2 Twoing Function.* The *Twoing* function operates by separating the classes into two groups. Say, below normal and normal classes belong to group one while the above normal class belongs to group two. *Twoing* then attempting to separate the two



groups in the descendant nodes of the tree. In essence, it's a measure of the difference in likelihood that a class appears in the left child node rather than the right child node. The objective is to make the likelihood that a given class case goes to the left as different as possible from the probability that it goes to the right. The function sums the absolute value of the probability differences over all classes. The formula used is:

$$(p_L \cdot p_R) \div 4 \sum_j (|p(j|t_L) - p(j|t_R)|^2), \quad (3)$$

where  $j$  and  $t$  represent a class and node, respectively (Brieman et al, 1984). The factor  $(p_L \cdot p_R)$  is the probability of a split left times the probability of a split right (designed to favor even splits). The term  $p(j|t_L)$  is a conditional probability statement. A similar method to the *Twoing* function is *Ordered Twoing*. In a multi-class node, such as the situation in this research, the class groups are restricted to consist of adjacent classes, such as below normal paired with normal or normal paired with above normal classes but not below normal paired with above normal. Overall, Brieman et al. (1984) noted that properties of the final tree were by and large insensitive to the choice of the impurity function, especially in the upper branches of the tree. Therefore, both the *Gini* and *Ordered Twoing* methods were used in this research.

*5.1.2 Priors.* To further assist the algorithm in making the best splits, the researcher must inform CART of the nature of the categorical class distribution. This is known as *Priors*. Simply put, knowledge gained independently of experience. In CART, there are several options. *Priors equal* means each class (below normal/normal/above normal) has an equal probability of occurring. *Priors data* means the probabilities of each class occurring match the total sample frequency (which is used in this research).

The remaining options are *priors test, learn, mix, and specified*; whose description are beyond the scope of this research.

*5.1.3 Improvement Score.* When a child node has a majority of observations that belong to one class, such as the below normal category, the node is considered more pure than the parent node it originated from, thus, a decrease in impurity has occurred. This is otherwise known as an improvement score, which measures how well the split improves the predictive performance of the tree. Subtracting from the parent node impurity, the sums of the child node impurities multiplied by their respective probabilities of a random case falling to the left or right child node, calculates the improvement score (Breiman et al., 1984). For example, if the impurity of a parent node was 0.626 (with 52 observations) and the left child node impurity was 0.408 (with only 7 observations) and the right child node impurity was 0.588 (with 45 observations) and the random case probabilities were  $7/52$  and  $45/52$ , respectively, the improvement score would be  $0.626 - (0.408 \times 7/52 + 0.588 \times 45/52)$  or 0.062.

*5.1.4 Pruning.* The splitting process repeats until it is not possible to split the node further or until a pre-specified child node size is attained. Breiman et al. (1984) recommended the approach of letting the splits continue until the terminal nodes are very small, as was done in this research. Without limiting the splits, eventually “pure” classification will be achieved. However, this “pure” result is usually unrealistic to follow as a conditional climatological tool, so this overgrown tree is then manually pruned upwards. Manual upward pruning of the tree results in terminal nodes with observation numbers that are large enough that they convey a sense of meaningful physical interpretation.

*5.1.5 Cross-validation.* With a large data set, classical regression often yields adequate test results. Ten-fold cross validation is valuable when no test sample is available and the learning sample is too small to have a test sample extracted from it. Ten-fold analysis grows the largest tree possible from the entire data set. Then ten random sub-samples, of roughly equal size are formed from this learning tree. The classification tree is computed ten times, each time leaving out one of the sub-samples from the computation. The omitted sub-sample is used as a test sample for the cross-validation. Thus, each learning sample is used nine times as a learning sample and once as a test sample. The error of all ten test samples is averaged to provide an overall estimate of the proportion of misclassified cases. The overall goal in selecting the optimal tree is to have a balance between the proportion of misclassified cases and an adequate tree size (avoiding a lengthy complex logical condition in the final decision tree).

## *5.2 CART Methodology and Results*

Prior to the construction of classification trees, the monthly OLR data sets for the country of Afghanistan (if the CART results were acceptable, the regional data sets would then be analyzed) had to be categorized to provide meaningful results. Each month, from September through May (from a climatological viewpoint, the cloudier months of the year) was standardized to make the distributions of values easy to compare against the mean. For example, the September mean was subtracted from the observed September value and then divided by the September standard deviation. After each month's data set was standardized, it was divided into three categories. Based on the

premise that OLR is strongly correlated in a negative manner to cloud cover, the middle 50% of the values was labeled as normal monthly cloudiness, the two remaining categories with proportions of 25% apiece, were labeled as above normal and below normal total monthly cloud cover. This division of the climatological categories was chosen based upon the relative usefulness of the predicted atmospheric variable for operational planners. Decision makers trying to plan for weather conditions one month away would be more interested in the extreme conditions of total monthly cloud cover. Thus, a 25-50-25 percent division was considered more relevant. Again, the climatological categories of below normal/normal/above normal monthly conditions were based on the 52-year POR data set.

After categorizing the data sets, the monthly classification trees were constructed. In order to determine if a tree was the optimal tree for basing a predictive algorithm on, certain criteria were considered:

1. general purity of the tree;
2. observation size of the terminal nodes; and,
3. misclassification rate, or error rate from the cross-validation test sample.

Since there were no benchmarks for acceptable rates of misclassification, reviewing previously published research was critical. Rodionov and Assel (2000) used a misclassification rate of 20% in their study of winter severity in the Great Lakes region and the relation to teleconnection pattern characteristics, essentially a nowcast. Burrows and Assel (1992) conducted a study-using CART for predicting on a daily time frame, spatially averaged ice-cover in the Great Lakes region using a 10-20% error rate.

Unfortunately, long range forecast studies using CART were not discovered in previous research.

However, Franz Baur (1951) warned the potential harm resulting from inaccurate long-range forecasts was greater the longer the time interval for which the prediction was created. Baur had two suggestions when employing statistics for long-range forecasts: (1) the statistical method should be guided by physical considerations, and (2) the need for better-than-chance statistical associations that guarantee a successful forecast with a probability greater than 92%. Taking these viewpoints into consideration, a misclassification rate of 25% in the test sample was determined as an adequate threshold for this research. A rate higher than 25% may render the classification tree unstable or unsuitable for forecasting purposes. Since the division of the 52-year data set is 25%/50%/25%, the low threshold for misclassification affords the user to focus on the extreme conditions of monthly cloudiness that might operationally impede a campaign or permit it to proceed.

*5.2.1 Classification Trees.* After the OLR data (for Afghanistan as a whole) for each month was standardized and categorized, fully-grown trees were constructed for the nine months using CART. All possible predictor variables were incorporated into the models. The trees were then pruned by selecting the most effective n-values for the parent and child nodes, consideration of the relative purity of the nodes, improvement scores at each split, and the overall risk estimate determined by the cross-validation.

An example of an overgrown classification tree is presented in Figure 18, which was constructed using October predictor variables compared to the categorized November OLR variable. The misclassification rate was 38% for this overgrown tree,

and the overgrown tree shows the improvement scores beneath each split. This tree has a rather lengthy logical condition statement with some weak improvement scores and poorly populated nodes. Thus, it would benefit from pruning (albeit, at a cost to the misclassification rate), specifically at nodes 4, 7, and 10.

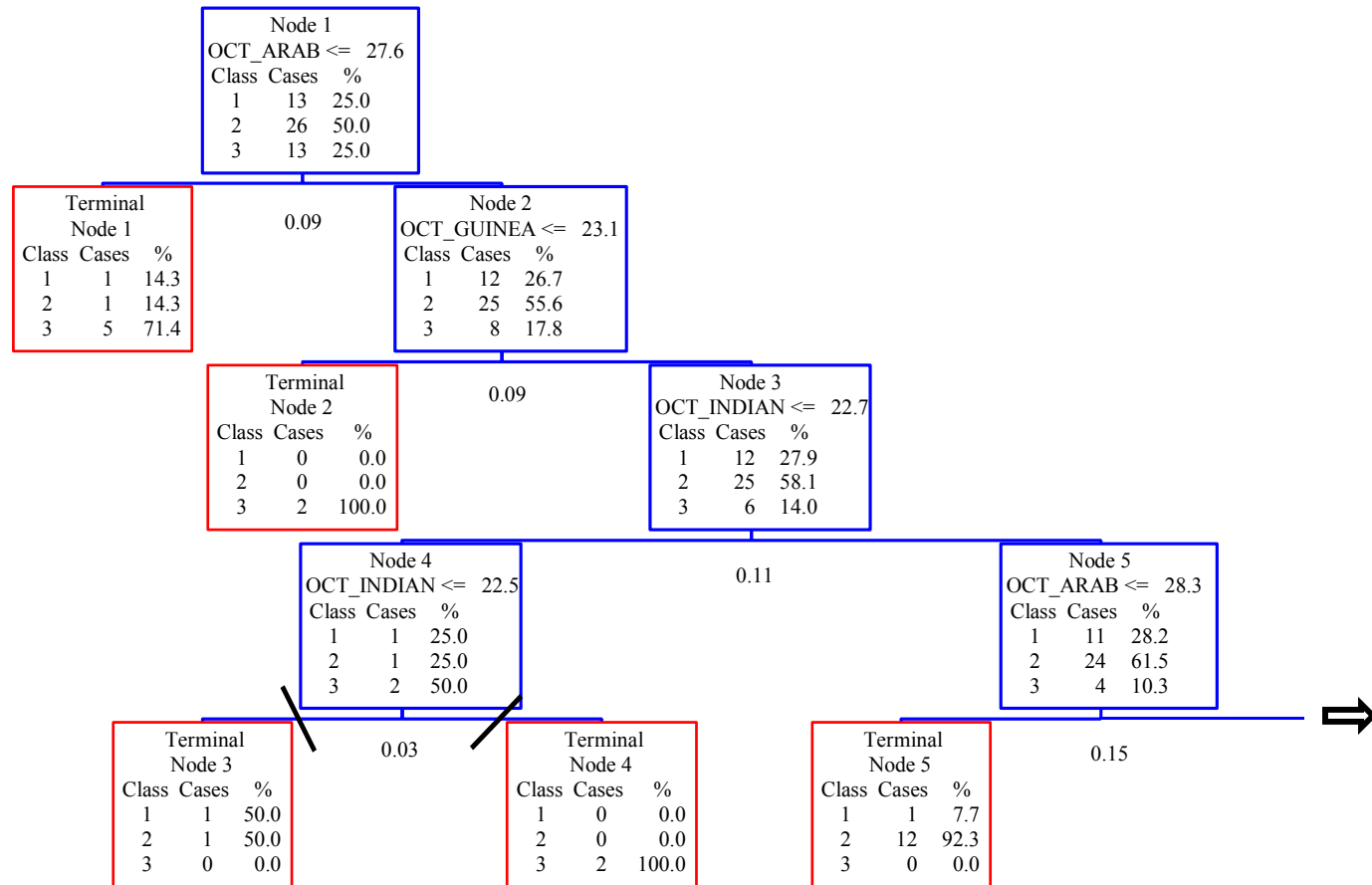


Figure 18a. An overgrown November tree. In Node 1, the second line contains the conditional statement. A yes response follows down the next level and into the left node, while a no response follows down a level and to the right. Slash marks indicate where pruning will occur. Arab = Arabian SST (°C); Guinea = Gulf of Guinea SST (°C); Indian = Indian Ocean SST (°C); Med = Mediterranean SST (°C); SCA = Scandinavian TI.

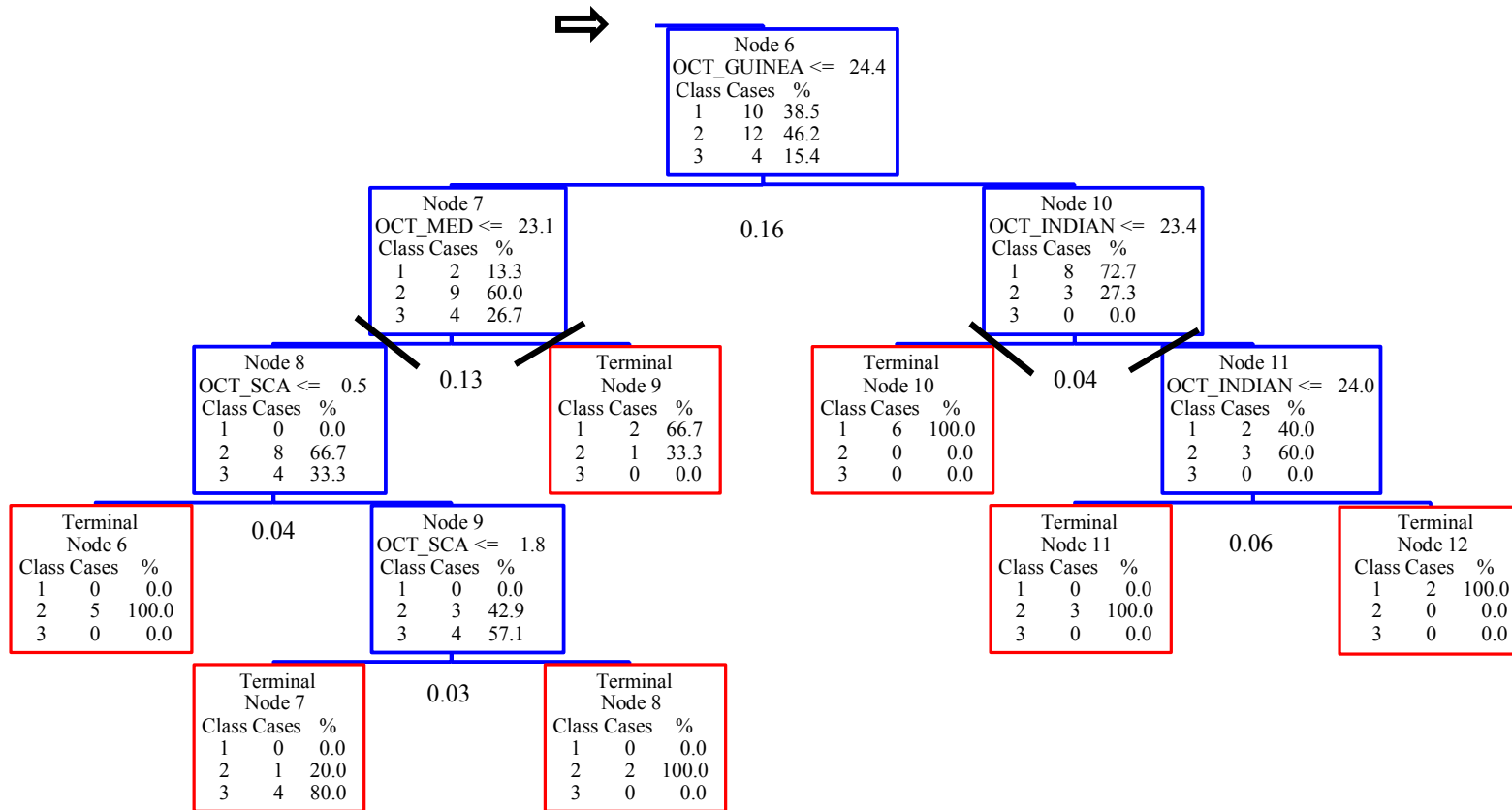


Figure 18b. November tree (continued) from Node 5.



Figure 19 shows the tree after the first pruning at Node 4 from Figure 18. Node 4 contained only four cases. When the splitting question of the October Indian SST being less than or equal to  $22.5^{\circ}\text{C}$  was implemented, CART sent two cases to the left child node and two to the right child node. As a result, the improvement score was 0.03 with one case populating Class 1 (below normal) and Class 2 (normal), respectively, in the left child node and two cases populating Class 3 (above normal) of the right child node. Since the cases were spread out amongst all three classes, and a weak improvement score resulted, it was deemed more prudent not to let further splitting occur below Node 4. Thus, Node 4 became Terminal Node 3 and would be considered an above normal monthly-cloudiness condition. As well, the misclassification rate has risen to 40%.

Figure 19b shows the location of the second pruning by CART to be at Node 6, where the question was asked whether the October Mediterranean SST was less than or equal to  $23.1^{\circ}\text{C}$ . To permit splitting below Node 6 would result in weak improvement scores of 0.04 and 0.03 for the first and second split, respectively. Thus, Node 6 would be classified as a normal monthly-cloudiness condition. The resulting tree can now be seen in Figure 20, and the misclassification rate using this tree would be 44%.

Figure 20b has the final pruning at Node 6. Although CART would split the remaining cases fairly well beyond Node 6, the case numbers would become rather small and it is more prudent to stop at Node 6 and classify the node as a below normal, monthly-cloudiness condition.

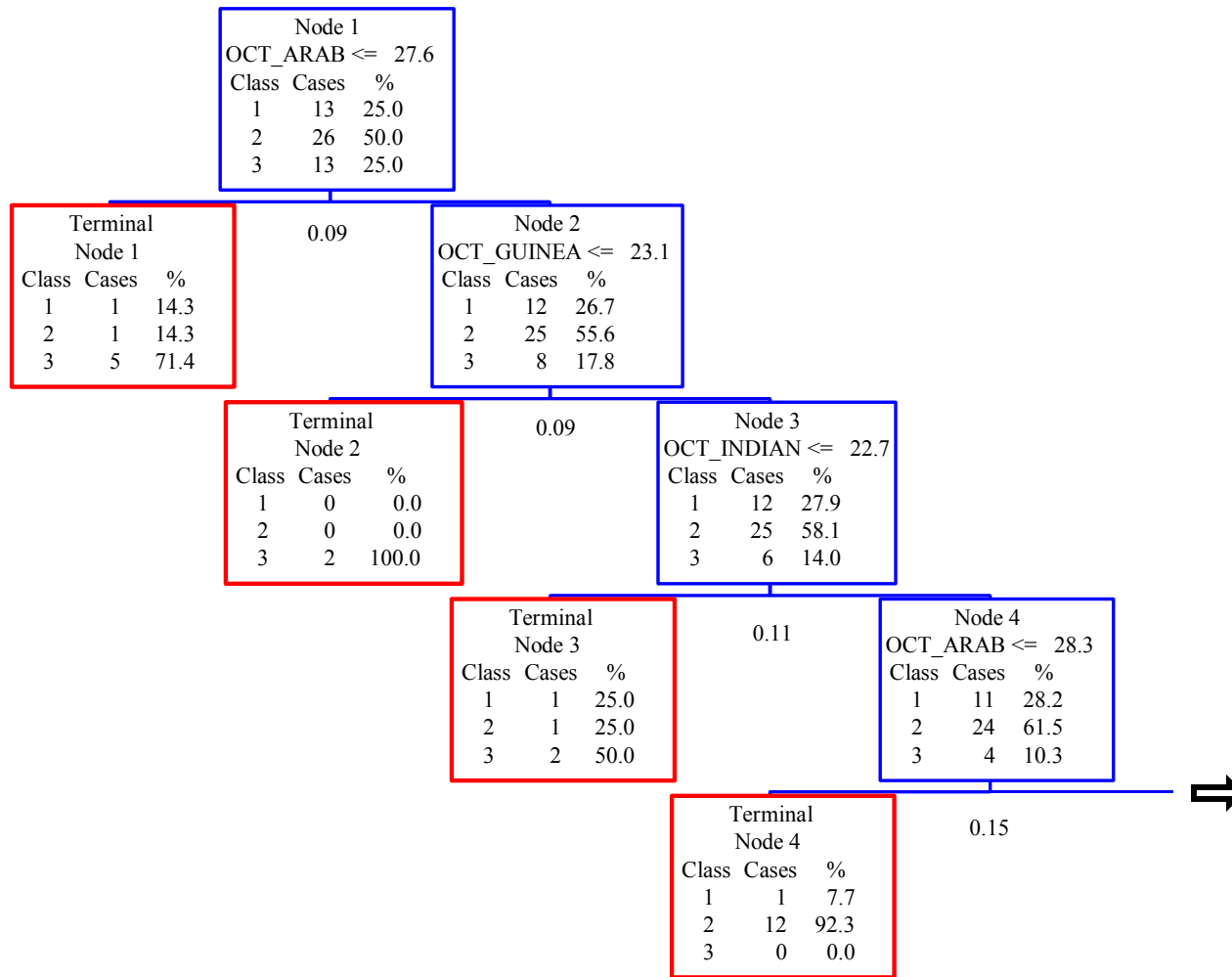


Figure 19a. November tree after first pruning.

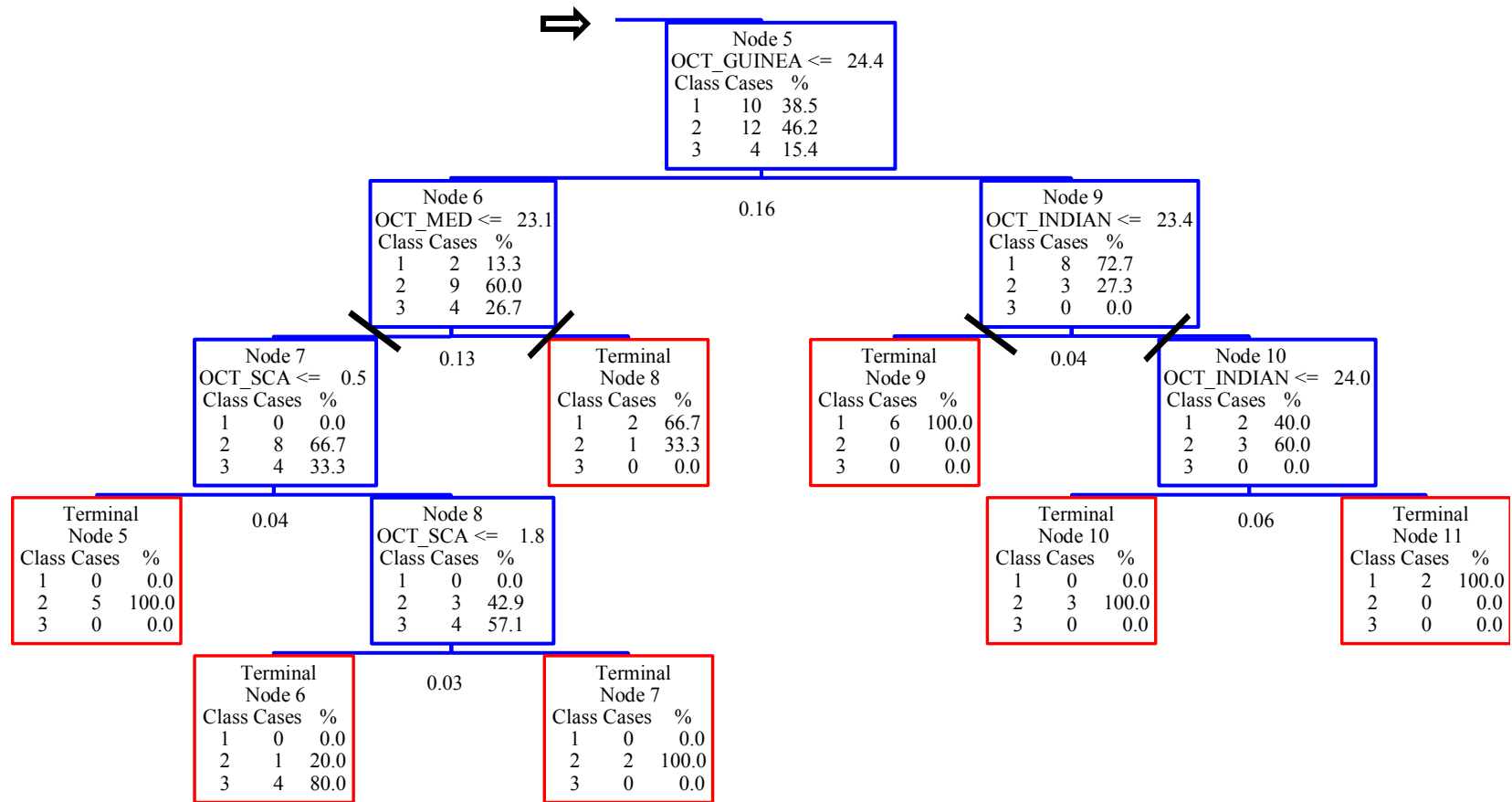


Figure 19b. First pruning of November tree (continued) from Node 4.

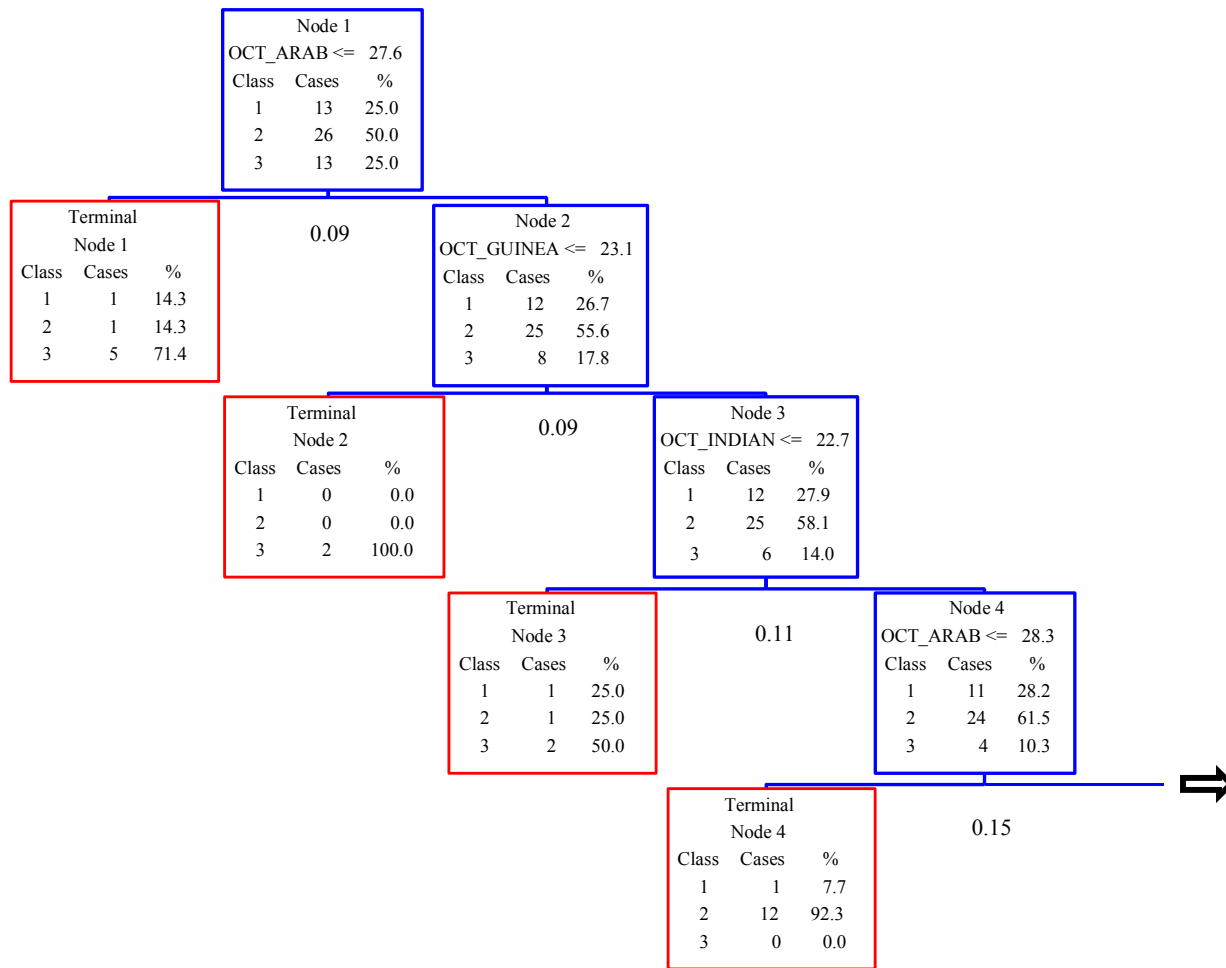


Figure 20a. November tree after second pruning.

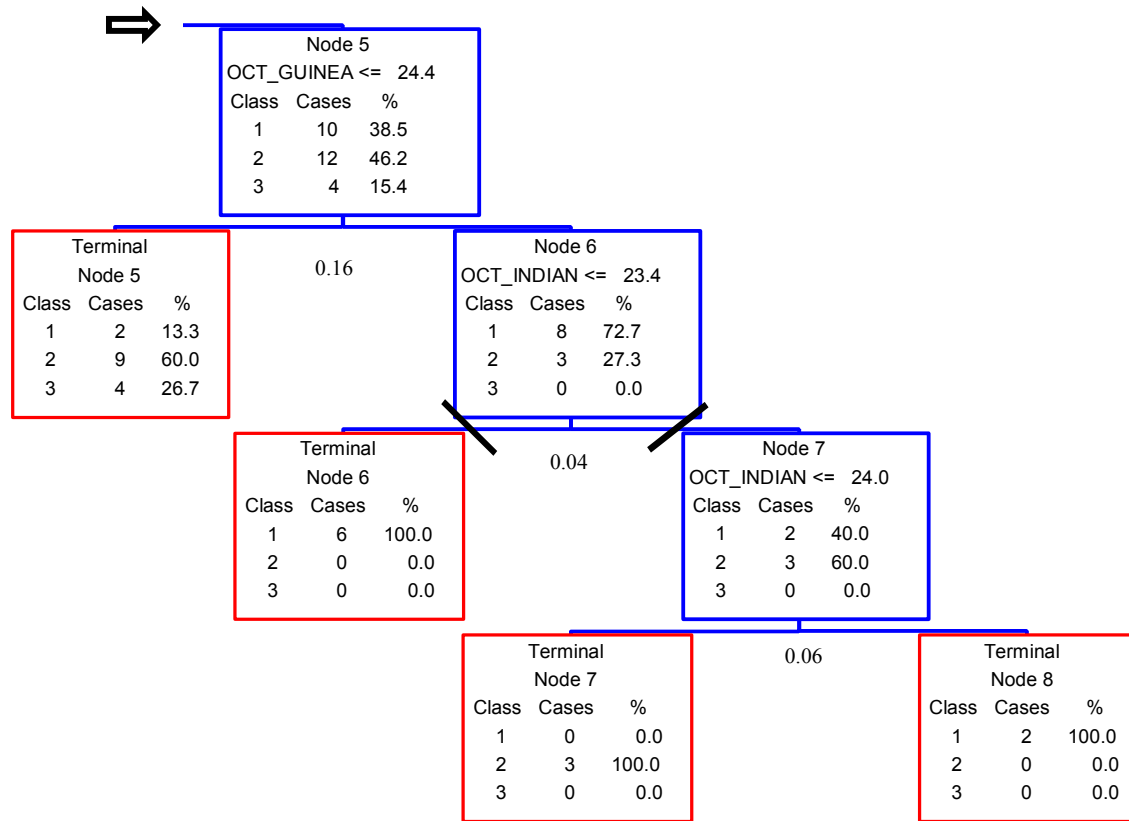


Figure 20b. November tree (continued) after second pruning.

Figure 21 shows the final tree with only five splitting levels. The overall misclassification rate is 48%, yet the logical condition is less complex than the original tree. To read the logical condition of the tree, one would begin as follows: If the October Arabian SST is less than or equal to  $27.6^{\circ}\text{C}$ , there is a 71% chance of above-normal total cloud cover for the month of November, otherwise a 29% chance exists of normal to below normal conditions (see terminal node 1). If the October Arabian SST is greater than  $27.6^{\circ}\text{C}$ , then proceed to the next level down the tree. If the October Guinea SST is less than or equal to  $23.1^{\circ}\text{C}$ , there is a very high likelihood of occurrence for above normal monthly-cloudiness. If the October Guinea SST is greater than  $23.1^{\circ}\text{C}$ , then proceed to the next level down the tree. If the Indian SST is less than or equal to  $22.7^{\circ}\text{C}$ , there is a 50% chance of above normal monthly-cloudiness, otherwise normal or below normal conditions may occur. If the Indian SST is greater than  $22.7^{\circ}\text{C}$ , then proceed to the next level down the tree. If the Arabian SST is less than or equal to  $28.3^{\circ}\text{C}$ , there is a 92% chance of normal monthly total cloud cover, otherwise, a nine percent chance exists of below normal conditions. If the October Arabian SST is greater than  $28.3^{\circ}\text{C}$ , then proceed to the next level down the tree. If the October Guinea SST is less than or equal to  $24.4^{\circ}\text{C}$ , a 60% chance of normal conditions may occur. If the October Guinea SST is greater than  $24.4^{\circ}\text{C}$ , there is a 73% chance of below normal monthly total cloud cover, otherwise, a 27% chance exists of normal total monthly cloud cover. This classification tree correctly classified 27 out of 52 cases. Again, the misclassification rate for this tree was 48%, well above the established threshold for extended range forecasting. Most of the occurrences of misclassification were spread out among all three classes, signifying high impurity in the nodes. The variables with the greatest contribution or importance to

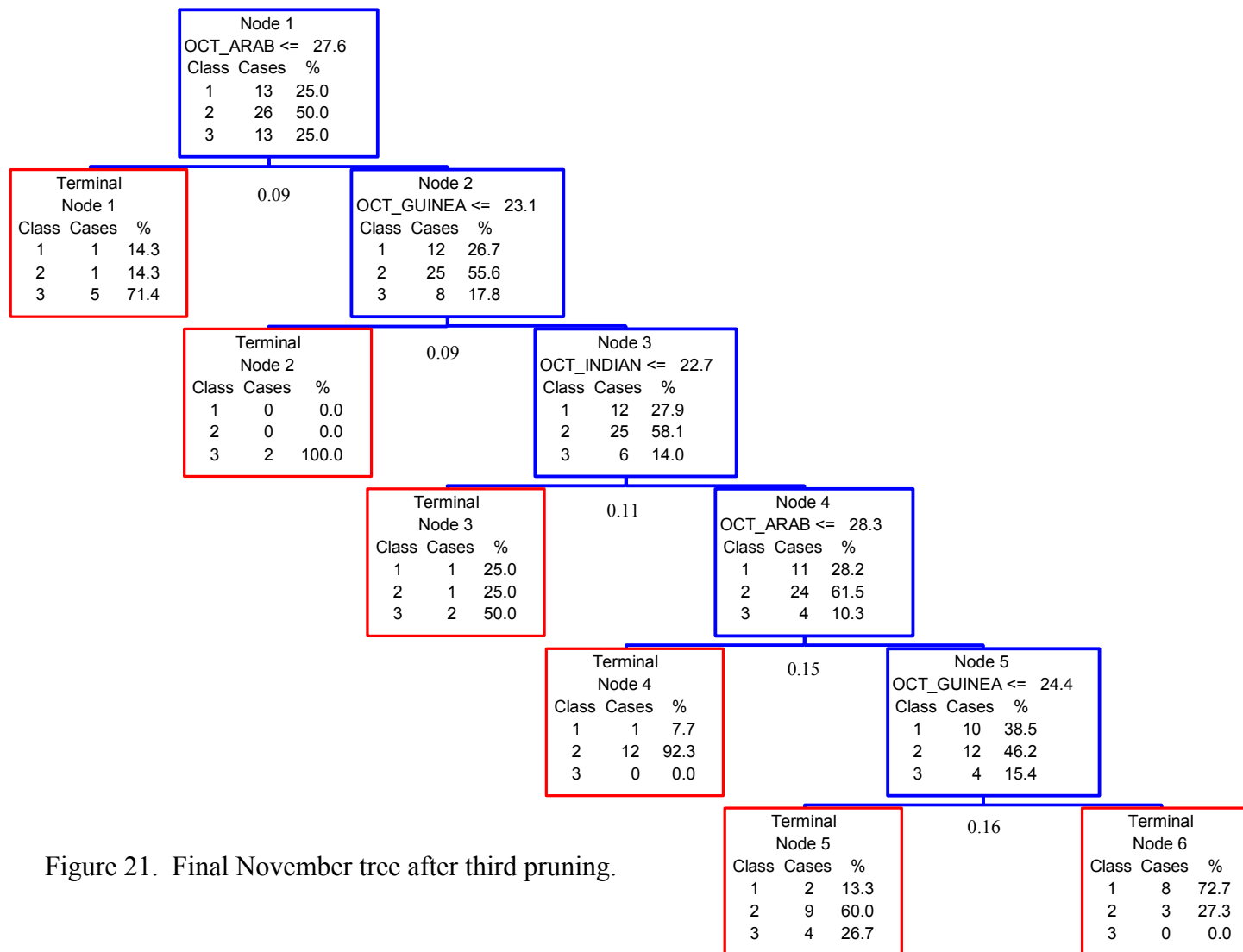


Figure 21. Final November tree after third pruning.

the tree are the Arabian and Guinea SSTs, as they have the higher improvement scores and are incorporated twice into the tree.

In this type of exploratory analysis, it is important to look at the data from various aspects. Just because a tree may not use all the predictors doesn't necessarily mean there's an insignificant association with the predictand. The reality may be that its effect was masked by other predictor variables. The overgrown November tree did incorporate a TI (Scandinavian pattern) as a splitting criteria, however, the improvement scores were less significant. When the manual pruning was completed, these nodes, which contained more impurity and smaller node cases, were dropped. The overgrown tree had a lower misclassification rate of 38% but it came at a cost of a lengthy, complex logical condition, which is not realistic to follow as a forecast tool. Had the misclassification rates met the threshold requirements, a final aspect to consider would be the improvement of the classification trees over the use of climatology, which is derived from the 52-year data set. Recall that the middle 50% of the standardized data was labeled as normal monthly total cloud cover, while the remaining two 25% proportions were labeled as above normal and below normal monthly total cloud cover, respectively.

Subtracting the climatological category value from the respective category value in the cross-validation test sample and then dividing by climatology would determine an improvement over climatology. For example: November's below normal, normal, and above normal categories had tree accuracy rates from the cross validation test of 38%, 69%, and 46%, in that order. First subtract 25%, 50%, and 25% from those tree accuracy rates, respectively. Then divide that difference by 25%, 50%, and 25% correspondingly, and the overall improvement scores are 52%, 38%, and 84%, respectively. For example,



when forecasting an above normal occurrence in November, the CART model would be 52% more likely to accurately forecast the event than climatological tables.

Figures 21-23 represent the pruned trees for the months of January, February, and March, in that order. The January tree focuses on just two predictors, the December Gulf of Guinea SST and the Scandinavian TI. In qualitative terms, this rule focuses on a narrow range of SST values and a TI pattern that is not substantially positive to make its best splits. However, the improvement scores are relatively low, and the misclassification rate of 48% is rather poor. The February tree also relies on two predictors, the January PNA TI and Mediterranean SST. The best split is the first split and the remaining splits are relatively poor. A warmer SST value and a slightly positive TI value leads to the best Class 1 (below normal) value of 80%. In general, the opposite requirements lead to Terminal Nodes 1 and 2, which have slightly favorable case numbers for the normal and above normal category. The March tree has a more complex logical condition based on the number of splits. CART uses two TIs and the Mediterranean SST to construct the tree. The best split again occurs at the top with an improvement score of 0.09 while the remaining splits are relatively low.

Table 5 lists the misclassification rates for each month from September to May as well as the improvement scores over climatology. Table 5 reveals that all the monthly classification trees exceeded the minimum threshold of a misclassification rate of 25%, as previously established to be of benefit for operational planning. One possible reason may be that the physical processes that cause cloud formation may not be adequately captured by the predictor variables. That leads to speculation there may exist other variables that could be incorporated into the CART analysis, which may increase the improvement

scores of the splits, leading to lower misclassification rates. As well, the effects of chaos may exist to the extent that long-range forecasting up to one month in advance is currently unattainable.

Patterns amongst the predictor variables within the classification trees were difficult to distinguish. Nevertheless, it was noted that overall, the pruned classification trees used SSTs as predictors in eight out of nine months. The Gulf of Guinea SST was used five times. The Mediterranean and Arabian SSTs were used 4 times each, and the Indian Ocean SST was used just once. TIs were incorporated by CART in only six of the nine months, with respect to the pruned trees.

The CART analysis focused on OLR over Afghanistan as a whole. In general, the classification tree results were not encouraging due to the high misclassification rates noted in the nine pruned trees. Thus, it was deemed unnecessary to proceed in individually analyzing the regions of the Northern Plains, Central Highlands, and Southwestern Lowlands in hopes of procuring adequate misclassification rates.

Table 5. Cross-validation misclassification rates of pruned classification trees and improvements over climatology for the months of September through May.

	Misclassification rate	Category	Improvement over climatology
September	37%	Below normal	16%
		Normal	92%
		Above normal	52%
October	46%	Below normal	52%
		Normal	16%
		Above normal	148%
November	40%	Below normal	52%
		Normal	38%
		Above normal	84%
December	40%	Below normal	no improvement
		Normal	92%
		Above normal	52%
January	48%	Below normal	no improvement
		Normal	38%
		Above normal	116%
February	40%	Below normal	116%
		Normal	30%
		Above normal	180%
March	44%	Below normal	52%
		Normal	52%
		Above normal	44%
April	54%	Below normal	no improvement
		Normal	26%
		Above normal	24%
May	37%	Below normal	84%
		Normal	30%
		Above normal	208%

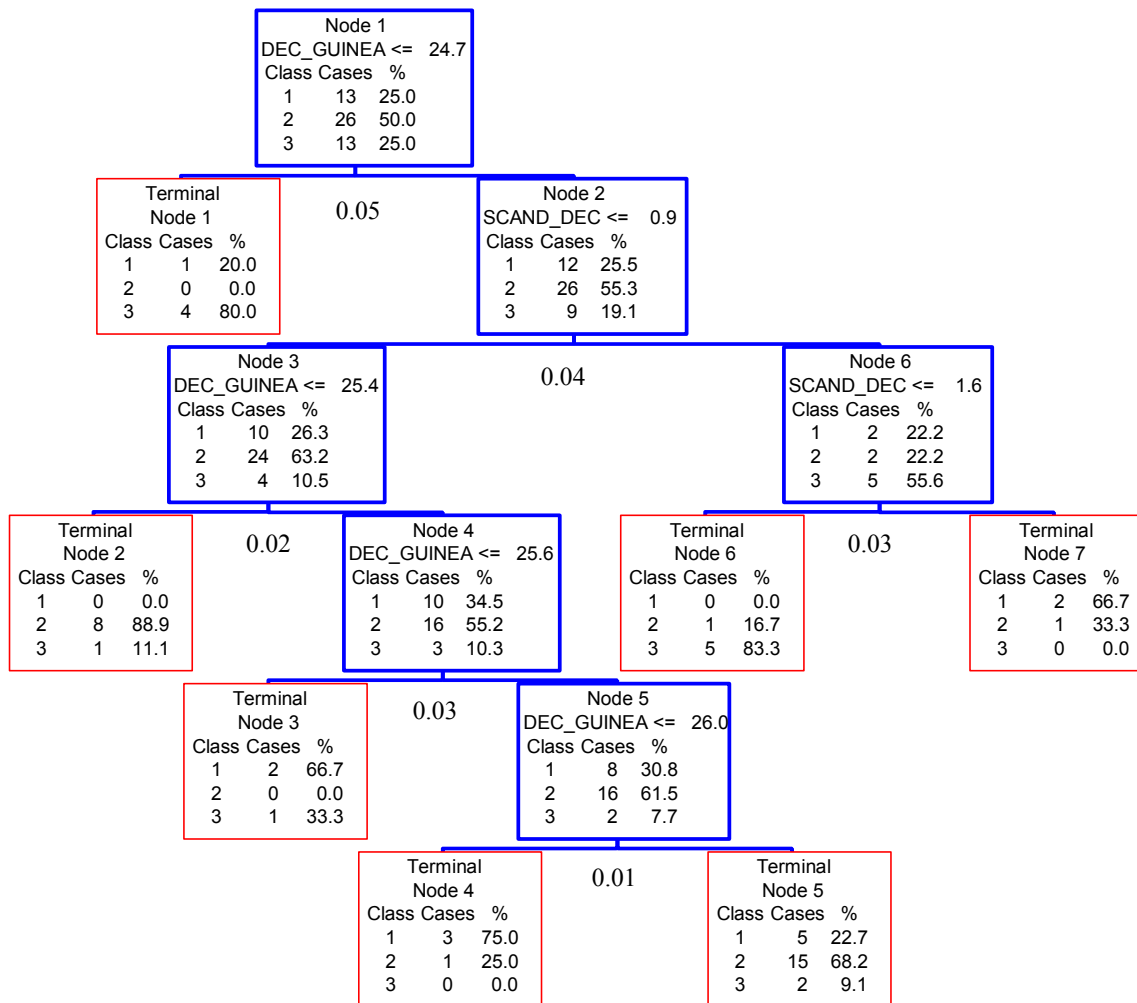


Figure 22. Pruned January classification tree with a misclassification rate of 48%.

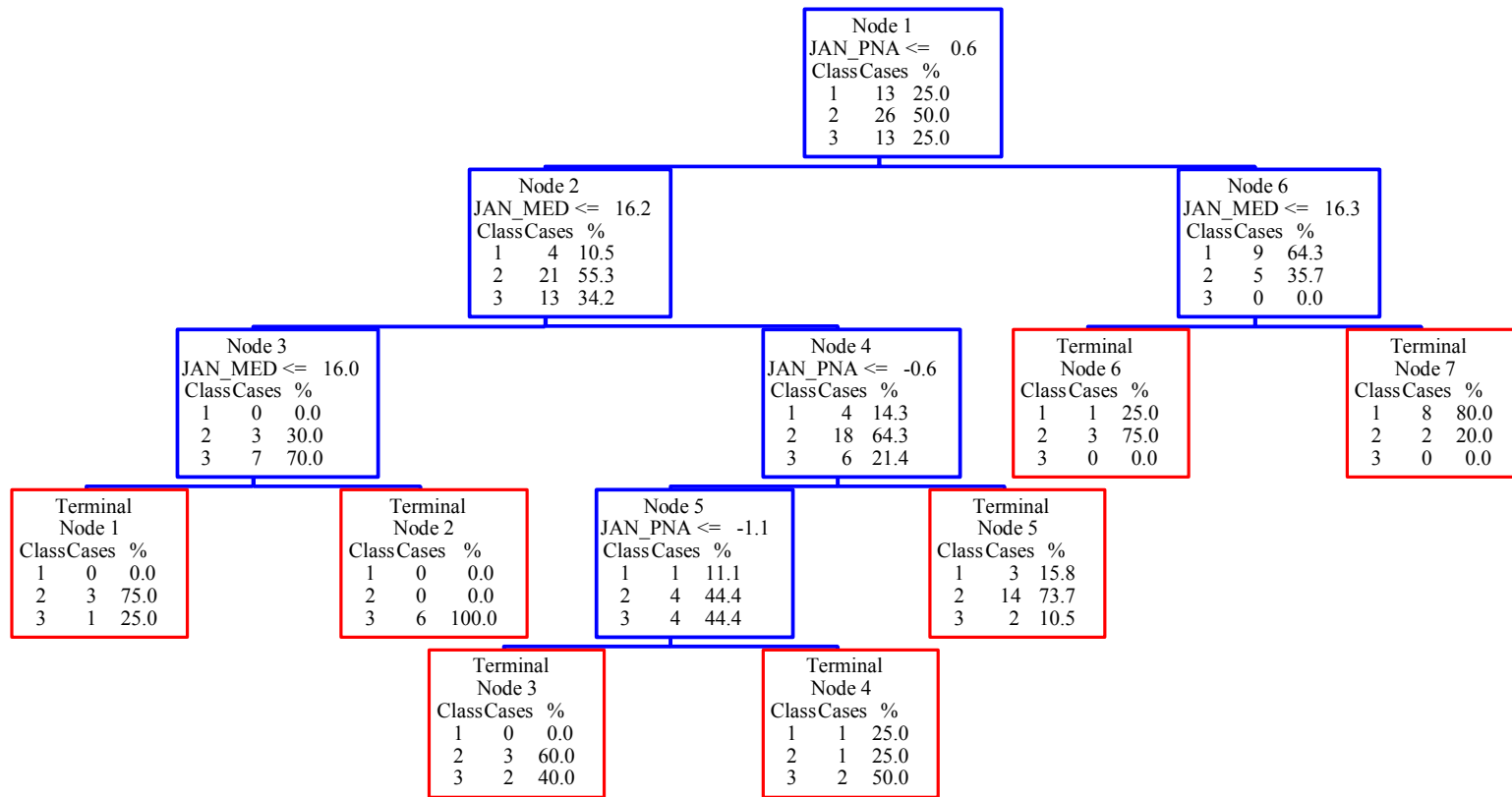


Figure 23. Pruned February tree with a misclassification rate of 40%.

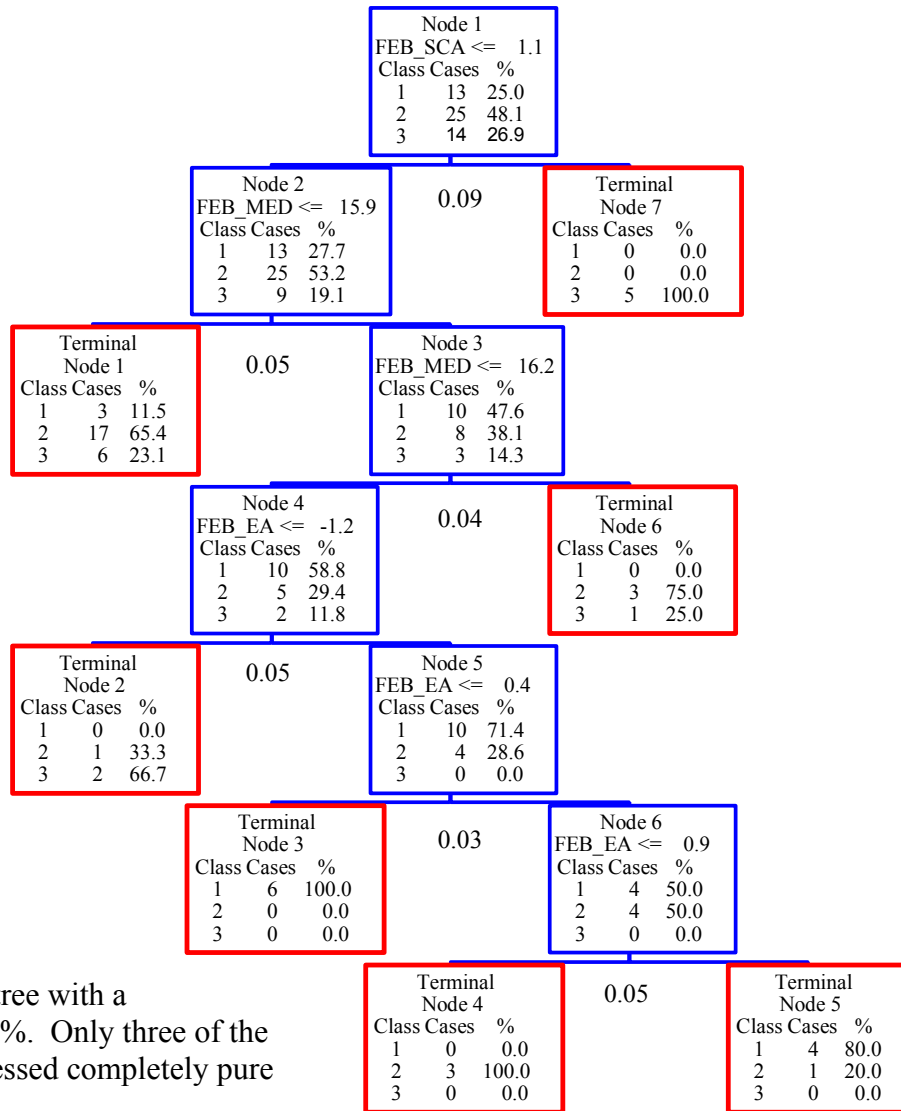


Figure 24. Pruned March tree with a misclassification rate of 44%. Only three of the seven terminal nodes possessed completely pure nodes.

## VI. Conclusions and Recommendations

### 6.1 Conclusions

The goal of this research was to construct a predictive tool for monthly total cloud cover over the country of Afghanistan. Standard statistical methodology was initially applied to search for predictive relationships between OLR and SSTs and TIs. As well, surface observational and RTNEPH data were considered as potential predictands. After extensive analysis using linear regression, CART was used to search for nonlinear relationships between the previously mentioned variables. Five of the six specific objectives, as stated in Chapter 1 were achieved.

The first objective consisted of gaining knowledge of the general climate of Afghanistan and conducting an overview of the synoptic conditions. Afghanistan possesses mainly arid to semi-arid climates. There are several regions: the Southwestern Lowlands, the Central Highlands (which make up the majority of the country), and the Northern Plains. Summers are relatively cloud-free due to the hot conditions brought about by the thermal low over southern Asia. In the fall months, the region experiences a change from the average pressure patterns of summer to winter, as the thermal low decreases in intensity due to the low angle of the sun as it retreats equatorward. Most extratropical cyclones pass well north of Afghanistan during a good portion of the year. During the winter, a semi-permanent high pressure region resides over the Asian continental interior bringing extremely cold temperatures to the region. This time of year has the greatest occurrence of cyclones transiting Afghanistan, bringing extensive cloud cover with them. The intensity and coverage of the high pressure system gradually

decreases, as well of the frequency of storm tracks, with the passage of the spring months and the onset of the thermal low that appears again over the southern portion of Asia in the summer months.

After gaining an understanding of the climate and synoptic systems that influence the region's weather, a collection of OLR, RTNEPH, surface observational data, teleconnection, and sea surface temperature data were assembled (objective 2). The data were based on monthly averages with the greatest attainable POR. The OLR data were also standardized and categorized as below normal/normal/above normal based on a distribution of 25%/50%/25%. This information was implemented in the standard statistical analyses, as well as the CART analyses (objectives 3 and 4).

The statistical analysis revealed that OLR and surface observational data had a relatively strong negative correlation. Taking into account the larger POR for OLR, it was decided the predictand of choice was OLR (as a surrogate for cloud cover). Unfortunately, in the regression analysis no significant relationships (yielding low  $R^2$  numbers) between OLR and the predictor variables were discovered (objective 3). As it turned out, seasonality held the largest amount of explained variance, thus data mining through CART analysis was undertaken as an additional exploratory tool.

The use of data mining through CART analysis was accomplished in order to extract any useful predictive information from the data that was not uncovered through standard statistical means (objective 5). Classification analysis was performed on the cloudier months of the year. The main findings of the CART research can be viewed in Table 5. Figures 17-23 were also shown to expose the reader to some of the pruned trees. The sequences of conditional climatological statements proved to have misclassification



rates that were well above the established threshold of 25%. With the given variables, acceptable predictability could not be achieved for forecasting extended-range total monthly cloud cover over Afghanistan.

## *6.2 Recommendations*

This research has shown the prospect of using sea surface temperatures and teleconnection indices to construct extended-range forecasts for monthly total cloud cover. It was revealed that CART data mining tools have the potential be used as forecast decision aids, when traditional statistical analysis fails to produce a predictive result. However, there is still a significant level of uncertainty in extended-range forecasting of monthly total cloud cover, by virtue of the high misclassification rates in the test sample cross validation.

Continued research on using CART data mining for atmospheric science applications should be pursued and the following are suggestions for future research endeavors:

1. Acquire larger POR source of either surface observational or RTNeph data to run more robust CART analyses for predicting total cloud cover.
2. Investigate possibilities of using other predictor variables in conjunction with SSTs and TIs to increase the potential for better improvement scores amongst the CART trees.
3. Consider the creation of a program to automatically ingest data via the internet to produce a cloud cover forecast from computed predictive decision aids.

## APPENDIX A: Correlogram

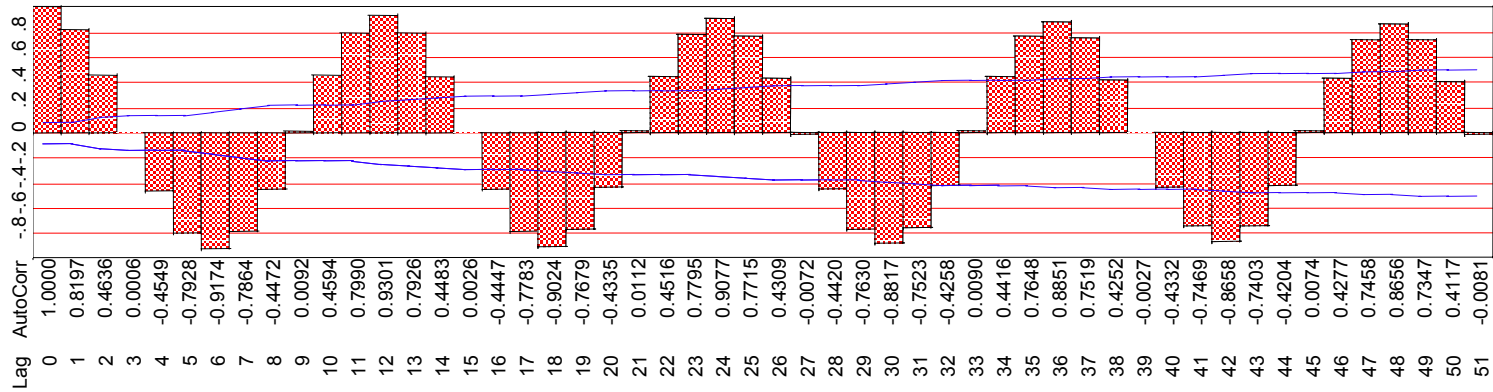


Figure A1. The above correlogram shows the correlation between the monthly Afghanistan OLR observations at different time distances apart. The solid lines represent  $\pm 2$  standard errors for approximately 95% confidence limits. This analysis shows OLR data points taken over time may have internal structure (such as autocorrelation or seasonal variation) that should be noted. Thus, it may adequately explain the magnitude of the  $R^2$  values recorded in Table 4.

## APPENDIX B: LINEAR REGRESSION RESULTS

Table 6. Linear regression results of TIs vs. predictands for the middle month of each season. NAO = North Atlantic Oscillation, POL = Polar/Eurasia, EP = East Pacific, EA/WR = East Atlantic/Western Russia, SCA = Scandinavia, ASU = Asian Summer.

	R <sup>2</sup>	P-value	RMSE
Jan OLR vs Dec NAO	0.00	0.96	0.98
Jan OLR vs Dec POL	0.01	0.42	0.97
Jan OLR vs Dec EA/WR	0.03	0.23	0.96
Jan OLR vs Dec SCA	0.03	0.27	0.97
Jan Obs vs Dec NAO	0.00	0.81	8.01
Jan Obs vs Dec POL	0.03	0.39	7.91
Jan Obs vs Dec EA/WR	0.12	0.07	7.53
Jan Obs vs Dec SCA	0.02	0.50	7.95
Jan RTNEPH vs Dec NAO	0.06	0.32	9.65
Jan RTNEPH vs Dec POL	0.33	0.01	8.18
Jan RTNEPH vs Dec EA/WR	0.00	0.82	9.95
Jan RTNEPH vs Dec SCA	0.00	0.92	9.96
Apr OLR vs Mar NAO	0.00	0.66	1.01
Apr OLR vs Mar EP	0.15	0.04	0.93
Apr OLR vs Mar EA/WR	0.10	0.03	0.96
Apr OLR vs Mar SCA	0.01	0.61	1.01
Apr Obs vs Mar NAO	0.00	0.79	7.69
Apr Obs vs Mar EP	0.12	0.06	7.21
Apr Obs vs Mar EA/WR	0.00	0.84	7.69
Apr Obs vs Mar SCA	0.04	0.32	7.56
Apr RTNEPH vs Mar NAO	0.08	0.26	8.56
Apr RTNEPH vs Mar EP	0.28	0.02	7.54
Apr RTNEPH vs Mar EA/WR	0.04	0.39	8.71
Apr RTNEPH vs Mar SCA	0.00	0.92	8.91
Jul OLR vs Jun ASU	0.13	0.008	0.94
Jul Obs vs Jun ASU	0.02	0.46	5.52
Jul RTNEPH vs Jun ASU	0.01	0.65	4.15
Oct OLR vs Sep NAO	0.02	0.36	1.00
Oct OLR vs Sep SCA	0.03	0.21	0.99
Oct Obs vs Sep NAO	0.12	0.07	8.68
Oct Obs vs Sep SCA	0.00	0.64	9.21
Oct RTNEPH vs Sep NAO	0.07	0.26	6.74
Oct RTNEPH vs Sep SCA	0.09	0.21	6.67

Table 7a. Linear regression results of SSTs vs. predictands. Med = Mediterranean Sea, Arab = Arabian Sea, Indian = Indian Ocean, Guinea = Gulf of Guinea. Only the middle month of each season was used.

	R <sup>2</sup>	P-value	RMSE
Jan OLR vs Dec Med	0.02	0.37	0.89
Jan OLR vs Dec Arab	0.01	0.43	0.89
Jan OLR vs Dec Indian	0.00	0.94	0.90
Jan OLR vs Dec Guinea	0.09	0.05	0.86
Jan Obs vs Dec Med	0.12	0.07	6.95
Jan Obs vs Dec Arab	0.00	7.40	0.80
Jan Obs vs Dec Indian	0.02	7.32	0.44
Jan Obs vs Dec Guinea	0.03	7.29	0.35
Jan RTNEPH vs Dec Med	0.00	0.79	9.94
Jan RTNEPH vs Dec Arab	0.03	0.46	9.80
Jan RTNEPH vs Dec Indian	0.16	0.10	9.12
Jan RTNEPH vs Dec Guinea	0.00	0.89	9.96
Apr OLR vs Mar Med	0.04	0.17	1.00
Apr OLR vs Mar Arab	0.04	0.15	1.00
Apr OLR vs Mar Indian	0.00	0.69	1.02
Apr OLR vs Mar Guinea	0.01	0.55	1.02
Apr Obs vs Mar Med	0.00	0.79	7.69
Apr Obs vs Mar Arab	0.02	0.46	7.62
Apr Obs vs Mar Indian	0.02	0.42	7.60
Apr Obs vs Mar Guinea	0.05	0.24	7.50
Apr RTNEPH vs Mar Med	0.01	0.72	8.88
Apr RTNEPH vs Mar Arab	0.00	0.84	8.91
Apr RTNEPH vs Mar Indian	0.00	0.91	8.91
Apr RTNEPH vs Mar Guinea	0.18	0.08	8.10

Table 7b. Regression results of SSTs vs. predictands for selected months.

	R <sup>2</sup>	P-value	RMSE
Jul OLR vs Jun Med	0.07	0.07	0.95
Jul OLR vs Jun Arab	0.05	0.12	0.96
Jul OLR vs Jun Indian	0.00	0.94	0.99
Jul OLR vs Jun Guinea	0.07	0.06	0.95
Jul Obs vs Jun Med	0.01	0.60	5.55
Jul Obs vs Jun Arab	0.00	0.90	5.58
Jul Obs vs Jun Indian	0.01	0.59	5.55
Jul Obs vs Jun Guinea	0.04	0.27	5.45
Jul RTNEPH vs Jun Med	0.34	0.01	3.38
Jul RTNEPH vs Jun Arab	0.09	0.22	3.99
Jul RTNEPH vs Jun Indian	0.13	0.13	3.89
Jul RTNEPH vs Jun Guinea	0.12	0.15	3.91
Oct OLR vs Sep Med	0.01	0.51	1.00
Oct OLR vs Sep Arab	0.01	0.43	1.00
Oct OLR vs Sep Indian	0.01	0.58	1.00
Oct OLR vs Sep Guinea	0.00	0.64	1.01
Oct Obs vs Sep Med	0.00	0.87	9.25
Oct Obs vs Sep Arab	0.18	0.03	8.26
Oct Obs vs Sep Indian	0.08	0.14	8.86
Oct Obs vs Sep Guinea	0.01	0.57	9.19
Oct RTNEPH vs Sep Med	0.01	0.66	6.96
Oct RTNEPH vs Sep Arab	0.00	0.94	7.01
Oct RTNEPH vs Sep Indian	0.00	0.84	7.00
Oct RTNEPH vs Sep Guinea	0.00	0.98	7.01

Table 8a. Multiple regression results for January OLR vs. December predictors.

	R <sup>2</sup>	P-value	RMSE
Jan OLR vs Dec Med & Dec NAO	0.02	0.65	0.90
Jan OLR vs Dec Med & Dec POL	0.03	0.48	0.90
Jan OLR vs Dec Med & Dec EA/WR	0.03	0.57	0.90
Jan OLR vs Dec Med & Dec SCA	0.03	0.47	0.90
Jan OLR vs Dec Arab & Dec NAO	0.01	0.73	0.91
Jan OLR vs Dec Arab & Dec POL	0.04	0.43	0.89
Jan OLR vs Dec Arab & Dec EA/WR	0.02	0.68	0.90
Jan OLR vs Dec Arab & Dec SCA	0.02	0.59	0.90
Jan OLR vs Dec Indian & Dec NAO	0.00	0.96	0.91
Jan OLR vs Dec Indian & Dec POL	0.02	0.65	0.90
Jan OLR vs Dec Indian & Dec EA/WR	0.00	0.93	0.91
Jan OLR vs Dec Indian & Dec SCA	0.01	0.80	0.91
Jan OLR vs Dec Guinea & Dec NAO	0.09	0.14	0.87
Jan OLR vs Dec Guinea & Dec POL	0.12	0.07	0.86
Jan OLR vs Dec Guinea & Dec EA/WR	0.10	0.10	0.86
Jan OLR vs Dec Guinea & Dec SCA	0.10	0.11	0.87

Table 8b. Multiple regression results for January surface observational data vs. December predictors.

	R <sup>2</sup>	P-value	RMSE
Jan Obs vs Dec Med & Dec NAO	0.12	0.20	7.09
Jan Obs vs Dec Med & Dec POL	0.15	0.13	6.97
Jan Obs vs Dec Med & Dec EA/WR	0.21	0.05	6.73
Jan Obs vs Dec Med & Dec SCA	0.13	0.19	7.07
Jan Obs vs Dec Arab & Dec NAO	0.00	0.96	7.54
Jan Obs vs Dec Arab & Dec POL	0.05	0.56	7.38
Jan Obs vs Dec Arab & Dec EA/WR	0.09	0.32	7.23
Jan Obs vs Dec Arab & Dec SCA	0.00	0.96	7.55
Jan Obs vs Dec Indian & Dec NAO	0.03	0.66	7.44
Jan Obs vs Dec Indian & Dec POL	0.06	0.46	7.33
Jan Obs vs Dec Indian & Dec EA/WR	0.12	0.21	7.10
Jan Obs vs Dec Indian & Dec SCA	0.02	0.73	7.47
Jan Obs vs Dec Guinea & Dec NAO	0.03	0.64	7.43
Jan Obs vs Dec Guinea & Dec POL	0.07	0.37	7.26
Jan Obs vs Dec Guinea & Dec EA/WR	0.10	0.29	7.20
Jan Obs vs Dec Guinea & Dec SCA	0.04	0.64	7.42

Table 8c. Multiple regression results for January RTNEPH data vs. December predictors.

	R <sup>2</sup>	P-value	RMSE
Jan RTNEPH vs Dec Med & Dec NAO	0.08	0.54	9.88
Jan RTNEPH vs Dec Med & Dec POL	0.33	0.05	8.45
Jan RTNEPH vs Dec Med & Dec EA/WR	0.01	0.95	10.25
Jan RTNEPH vs Dec Med & Dec SCA	0.01	0.95	10.25
Jan RTNEPH vs Dec Arab & Dec NAO	0.08	0.54	9.88
Jan RTNEPH vs Dec Arab & Dec POL	0.33	0.05	8.43
Jan RTNEPH vs Dec Arab & Dec EA/WR	0.04	0.75	10.10
Jan RTNEPH vs Dec Arab & Dec SCA	0.03	0.77	10.10
Jan RTNEPH vs Dec Indian & Dec NAO	0.16	0.27	9.42
Jan RTNEPH vs Dec Indian & Dec POL	0.44	0.01	7.68
Jan RTNEPH vs Dec Indian & Dec EA/WR	0.17	0.26	9.40
Jan RTNEPH vs Dec Indian & Dec SCA	0.18	0.23	9.32
Jan RTNEPH vs Dec Guinea & Dec NAO	0.06	0.62	10.00
Jan RTNEPH vs Dec Guinea & Dec POL	0.33	0.05	8.45
Jan RTNEPH vs Dec Guinea & Dec EA/WR	0.01	0.95	10.26
Jan RTNEPH vs Dec Guinea & Dec SCA	0.00	0.99	10.28



Table 9a. Multiple regression results for April OLR vs. March predictors.

	R <sup>2</sup>	P-value	RMSE
Apr OLR vs Mar Med & Mar NAO	0.04	0.40	1.01
Apr OLR vs Mar & Mar EP	0.18	0.01	0.94
Apr OLR vs Mar Med & Mar EA/WR	0.13	0.05	0.97
Apr OLR vs Mar Med & Mar SCA	0.06	0.25	1.01
Apr OLR vs Mar Arab & Mar NAO	0.05	0.35	1.01
Apr OLR vs Mar Arab & Mar EP	0.18	0.01	0.95
Apr OLR vs Mar Arab & Mar EA/WR	0.13	0.04	0.97
Apr OLR vs Mar Arab & Mar SCA	0.05	0.33	1.01
Apr OLR vs Mar Indian & Mar NAO	0.01	0.88	1.04
Apr OLR vs Mar Indian & Mar EP	0.15	0.03	0.96
Apr OLR vs Mar Indian & Mar EA/WR	0.11	0.08	0.98
Apr OLR vs Mar Indian & Mar SCA	0.01	0.78	1.03
Apr OLR vs Mar Guinea & Mar NAO	0.01	0.73	1.03
Apr OLR vs Mar Guinea & Mar EP	0.16	0.02	0.95
Apr OLR vs Mar Guinea & Mar EA/WR	0.12	0.06	0.97
Apr OLR vs Mar Guinea & Mar SCA	0.02	0.69	1.03

Table 9b. Multiple regression results for April surface observational data vs. March predictors.

	R <sup>2</sup>	P-value	RMSE
Apr Obs vs Mar Med & Mar NAO	0.01	0.93	7.83
Apr Obs vs Mar & Mar EP	0.12	0.18	7.34
Apr Obs vs Mar Med & Mar EA/WR	0.00	0.95	7.83
Apr Obs vs Mar Med & Mar SCA	0.04	0.62	7.70
Apr Obs vs Mar Arab & Mar NAO	0.02	0.77	7.77
Apr Obs vs Mar Arab & Mar EP	0.15	0.13	7.25
Apr Obs vs Mar Arab & Mar EA/WR	0.02	0.75	7.76
Apr Obs vs Mar Arab & Mar SCA	0.07	0.37	7.56
Apr Obs vs Mar Indian & Mar NAO	0.02	0.72	7.75
Apr Obs vs Mar Indian & Mar EP	0.17	0.08	7.14
Apr Obs vs Mar Indian & Mar EA/WR	0.02	0.72	7.75
Apr Obs vs Mar Indian & Mar SCA	0.07	0.39	7.57
Apr Obs vs Mar Guinea & Mar NAO	0.07	0.41	7.58
Apr Obs vs Mar Guinea & Mar EP	0.15	0.12	7.23
Apr Obs vs Mar Guinea & Mar EA/WR	0.06	0.45	7.60
Apr Obs vs Mar Guinea & Mar SCA	0.09	0.30	7.50

Table 9c. Multiple regression results for April RTNEPH data vs. March predictors.

	R <sup>2</sup>	P-value	RMSE
Apr RTNEPH vs Mar Med	0.01	0.72	8.88
Apr RTNEPH vs Mar Arab	0.00	0.84	8.91
Apr RTNEPH vs Mar Indian	0.00	0.91	8.91
Apr RTNEPH vs Mar Guinea	0.18	0.08	8.10
Apr RTNEPH vs Mar Med & Mar NAO	0.08	0.53	8.83
Apr RTNEPH vs Mar Med & Mar EP	0.29	0.08	7.78
Apr RTNEPH vs Mar Med & Mar EA/WR	0.05	0.68	8.97
Apr RTNEPH vs Mar Med & Mar SCA	0.01	0.93	9.16
Apr RTNEPH vs Mar Arab & Mar NAO	0.09	0.51	8.80
Apr RTNEPH vs Mar Arab & Mar EP	0.29	0.08	7.78
Apr RTNEPH vs Mar Arab & Mar EA/WR	0.05	0.68	8.98
Apr RTNEPH vs Mar Arab & Mar SCA	0.00	0.97	9.19
Apr RTNEPH vs Mar Indian & Mar NAO	0.09	0.50	8.80
Apr RTNEPH vs Mar Indian & Mar EP	0.30	0.07	7.69
Apr RTNEPH vs Mar Indian & Mar EA/WR	0.05	0.70	9.00
Apr RTNEPH vs Mar Indian & Mar SCA	0.00	0.98	9.20
Apr RTNEPH vs Mar Guinea & Mar NAO	0.24	0.13	8.05
Apr RTNEPH vs Mar Guinea & Mar EP	0.39	0.02	7.17
Apr RTNEPH vs Mar Guinea & Mar EA/WR	0.37	0.03	7.31
Apr RTNEPH vs Mar Guinea & Mar SCA	0.18	0.22	8.32

Table 10. Multiple regression results for July predictands vs June predictors.

	R <sup>2</sup>	P-value	RMSE
Jul OLR vs Jun Med & Jun ASU	0.20	0.01	0.89
Jul OLR vs Jun Arab & Jun ASU	0.16	0.02	0.91
Jul OLR vs Jun Indian & Jun ASU	0.15	0.02	0.92
Jul OLR vs Jun Guinea & Jun ASU	0.20	0.01	0.89
Jul Obs vs Jun Med & Jun ASU	0.04	0.61	5.58
Jul Obs vs Jun Arab & Jun ASU	0.04	0.63	5.87
Jul Obs vs Jun Indian & Jun ASU	0.04	0.63	5.89
Jul Obs vs Jun Guinea & Jun ASU	0.09	0.31	5.44
Jul RTNEPH vs Jun Med & Jun ASU	0.36	0.04	3.46
Jul RTNEPH vs Jun Arab & Jun ASU	0.09	0.48	4.11
Jul RTNEPH vs Jun Indian & Jun ASU	0.15	0.29	0.04
Jul RTNEPH vs Jun Guinea & Jun ASU	0.12	0.37	4.04

Table 11. Multiple regression results for October predictands vs September predictors.

	R <sup>2</sup>	P-value	RMSE
Oct OLR vs Sep Med & Sep NAO	0.02	0.36	1.00
Oct OLR vs Sep Med & Sep SCA	0.03	0.48	1.00
Oct OLR vs Sep Arab & Sep NAO	0.02	0.57	1.01
Oct OLR vs Sep Arab & Sep SCA	0.04	0.40	1.00
Oct OLR vs Sep Indian & Sep NAO	0.02	0.71	1.01
Oct OLR vs Sep Indian & Sep SCA	0.03	0.49	1.00
Oct OLR vs Sep Guinea & Sep NAO	0.02	0.67	1.01
Oct OLR vs Sep Guinea & Sep SCA	0.03	0.49	1.00
Oct Obs vs Sep Med & Sep NAO	0.12	0.20	8.46
Oct Obs vs Dec Med & Sep SCA	0.01	0.87	9.38
Oct Obs vs Sep Arab & Sep NAO	0.28	0.02	8.01
Oct Obs vs Sep Arab & Sep SCA	0.18	0.08	8.53
Oct Obs vs Sep Indian & Sep NAO	0.19	0.08	8.51
Oct Obs vs Sep Indian & Sep SCA	0.09	0.32	9.02
Oct Obs vs Sep Guinea & Sep NAO	0.14	0.15	8.75
Oct Obs vs Sep Guinea & Sep SCA	0.02	0.79	9.35
Oct RTNEPH vs Sep Med & Sep NAO	0.09	0.50	6.92
Oct RTNEPH vs Dec Med & Sep SCA	0.10	0.47	6.89
Oct RTNEPH vs Sep Arab & Sep NAO	0.08	0.54	6.95
Oct RTNEPH vs Sep Arab & Sep SCA	0.10	0.46	6.88
Oct RTNEPH vs Sep Indian & Sep NAO	0.08	0.52	6.93
Oct RTNEPH vs Sep Indian & Sep SCA	0.10	0.45	6.87
Oct RTNEPH vs Sep Guinea & Sep NAO	0.08	0.53	6.94
Oct RTNEPH vs Sep Guinea & Sep SCA	0.10	0.46	6.88

### APPENDIX C: CART RESULTS FOR BAGHDAD, IRAQ

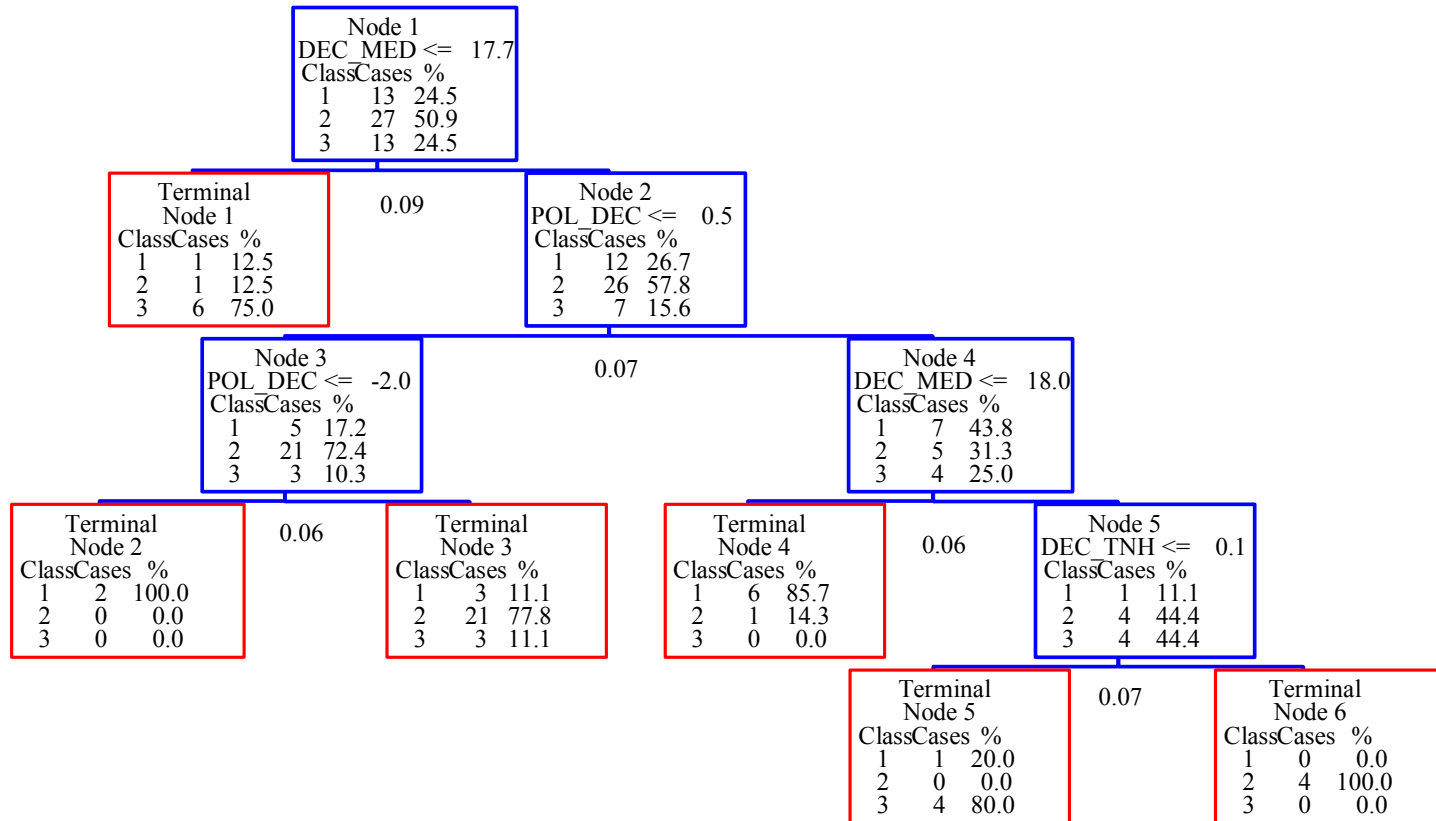


Figure C1. This January classification tree for Baghdad, Iraq OLR had a misclassification rate of 36%. Terminal nodes 2,4, and 6 were purer nodes considering the resultant cases where in adjacent nodes or the nodes consisted of only one class.

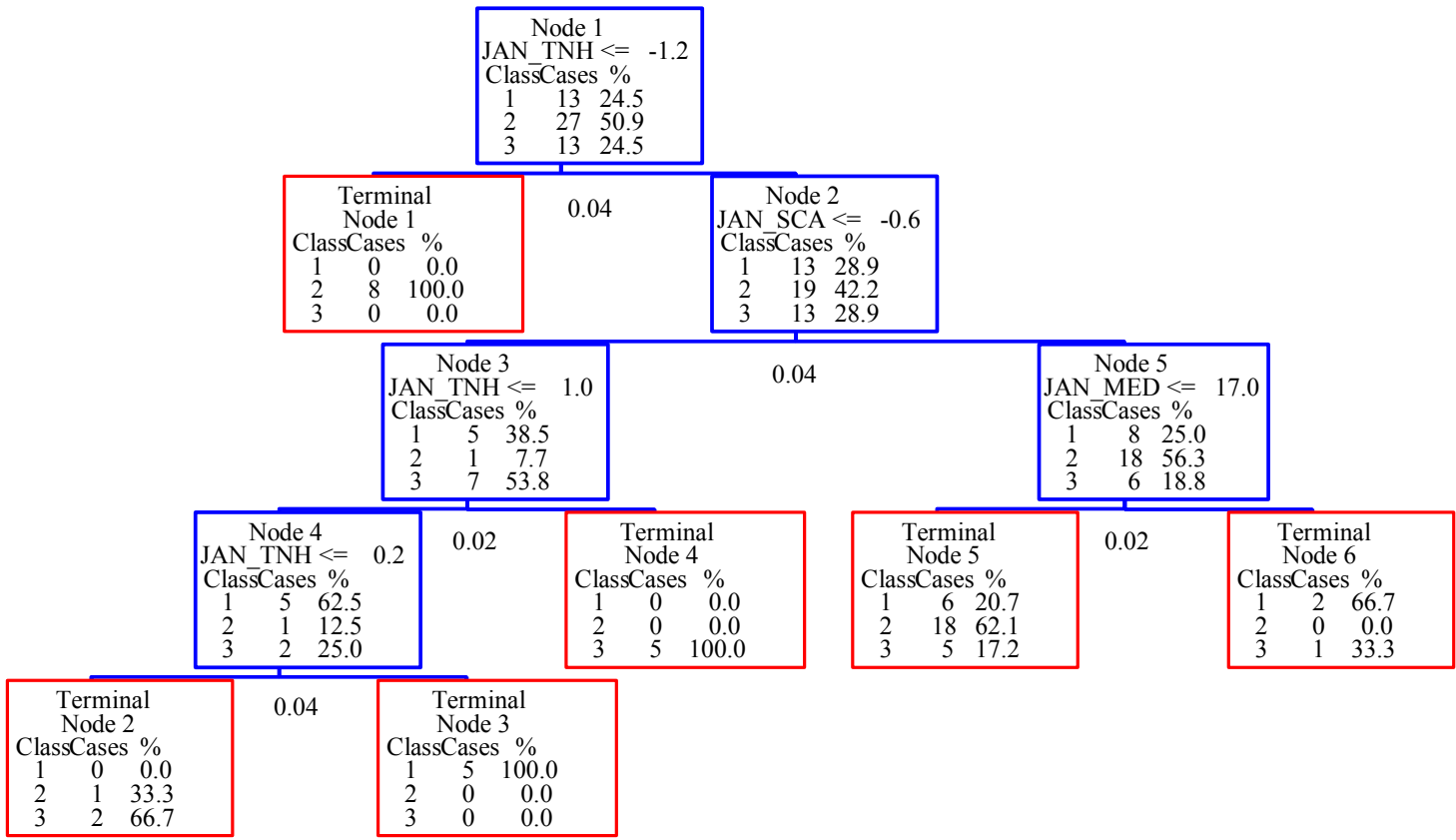


Figure C2. The February classification tree for Baghdad, Iraq OLR had a misclassification rate of 45%. The cross validation test sample had tree accuracy rates for Class 1,2, and 3 of 23%, 81%, and 31%, respectively.

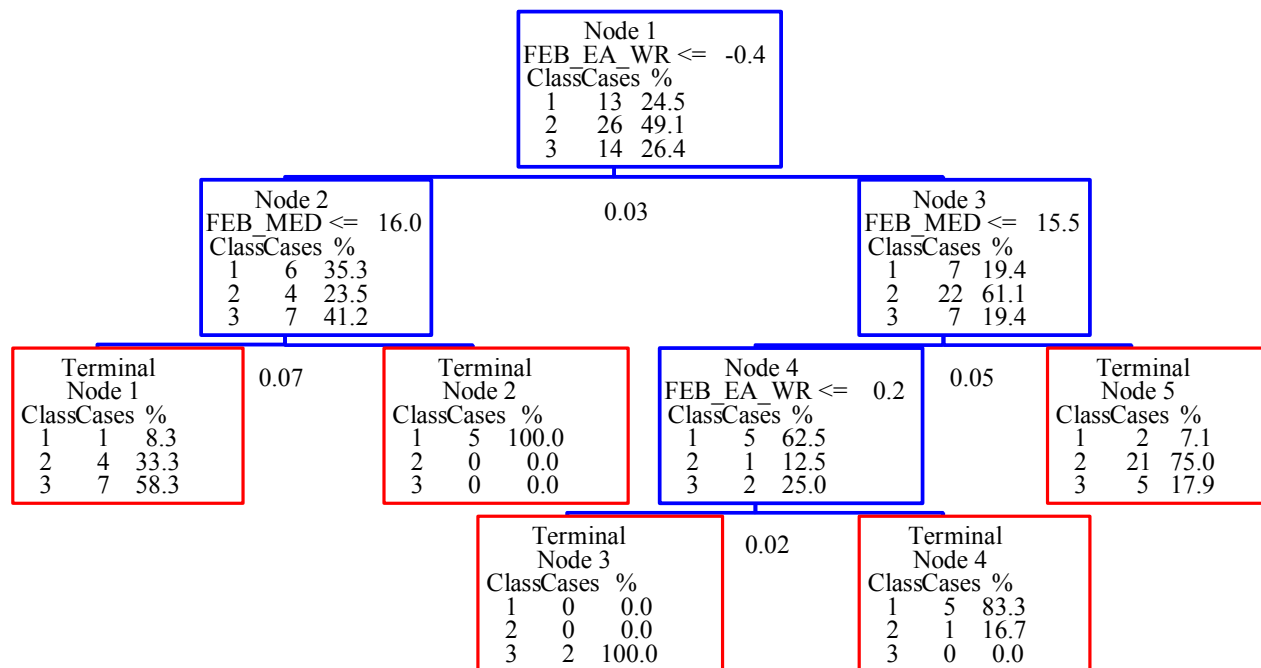


Figure C3. The March classification tree for Baghdad, Iraq OLR had a misclassification rate of 40%. The cross validation test sample had tree accuracy rates for Class 1,2, and 3 of 62%, 77%, and 29%, respectively.



## Bibliography

- AFCCC/DOC3/NCOIC. Asheville North Carolina. Personal Correspondence. 3 September 2002.
- Assel, Raymond A. and Burrows, William R. "Use of CART for Diagnostic and Prediction Problems in the Atmospheric Sciences," Proceedings of the 12<sup>th</sup> Conference on Probability and Statistics in the Atmospheric Sciences. 161-166. Boston: American Meteorological Society, 1992.
- Barnston, Anthony G. and Livezey, Robert E., 1986: *Classification, Seasonality and Persistence of Low-Frequency Atmospheric Circulation Patterns*, American Meteorological Society: Monthly Weather Review, **115**, 1083-1126.
- Baur, F., 1951: Extended-Range Weather Forecasting, *Compendium of Meteorology*, T.F. Malone, Ed., American Meteorology Society, 814-833.
- Bieker, Francis, 2002: Background Paper on Cloud Depiction and Forecast System II (CDFS II), 2pp.
- Breiman, L.; Friedman, J.H., Olshen, R.A. and Stone, C.J., 1984: *Classification and Regression Trees*. Wadsworth, Belmont, CA, 358pp.
- Carleton, Andrew M., 1991: *Satellite Remote Sensing in Climatology*, Belhaven Press, Boston, 291pp.
- Della-Rose, Dennis, Class lecture, METG 610, Radiative Transfer. School of Engineering Physics, Air Force Institute of Technology, Wright-Patterson AFB OH, Winter 2002.
- FLENUMMETOC DET, cited 2002: Surface Climatic Summaries. [Available online at <http://navy.ncdc.noaa.gov/products/online.html>.]
- Freestrom, Hugh J. "Designing an Algorithm to Predict the Intensity of the Severe Weather Season," Master's Thesis, Air Force Institute of Technology, Department of Engineering Physics, 4-6, 2002.
- Glantz, M.H., 1991: Introduction, *Teleconnections Linking Worldwide Climate Anamolies*, M.H. Glantz, R.W. Katz, & N. Nicholls, Ed., Cambridge University Press, New York, 1-11.
- Hartmann, Dennis L., 1994: *Global Physical Climatology*, Academic Press, New York, 411pp.
- Jenne, Roy. Data Support Section Manager, NCEP/NCAR, Boulder CO. Telephone interview. 2 October 2002.

- Jury, Mark R., 1995: *Regional Teleconnection Pattern Associated with Summer Rainfall over South Africa, Namibia, and Zimbabwe*, International Journal of Climatology: Royal Meteorological Society, **16**, 135-153.
- Kalnay, E. et al. 1996: *The NCEP/NCAR 40-Year Reanalysis Project*, American Meteorological Society, 1-47.
- Kiess, R.T. and W.M. Cox, *The AFGWC Automated Real-Time Cloud Analysis Model*. AFGWC/TN-88/001, AFGWC, Air Weather Service, Offutt AFB, NE, 1988.
- Lowther Ronald P. et al. 1991: RTNEPH Total Cloud Cover: Validation Study, USAFETAC/PR-91/020, Scott AFB, IL, 51pp.
- Makridakis, Spyros et al. *Forecasting: Methods and Applications*. New York: John Wiley & Sons, 1998.
- Merriam and Webster, 2001: *Collegiate Dictionary*. 10<sup>th</sup> ed. Merriam-Webster Incorporated, 1557pp.
- National Geographics, cited 2002: Expeditions Atlas. [Available online at <http://nationalgeographics.com/xpeditions/atlas/>.]
- National Intelligence Survey, 1970: Iran and Afghanistan: Weather and Climate, Notice number 131, NIS, 179pp.
- Randall, Robb M. "Exploration of Teleconnection Indices for Long-Range Seasonal Temperature Forecasts," Master's Thesis, Air Force Institute of Technology, Department of Engineering Physics, 1-5, 2002.
- Raval, Ramaswany and A.H. Oort, 1994: *Observed Dependence of Outgoing Longwave Radiation on Sea Surface Temperature and Moisture*, American Meteorological Society: Journal of Climate, **7**, 807-821.
- Rodionov, Sergei and Assel Raymond, 2000: Atmospheric Teleconnection Patterns and Severity of Winters in the Laurentian Great Lakes Basin, NOAA: Atmosphere-Ocean, **38**, 601-635.
- SAS Institute Inc., cited 2002: Help menu. [Available online at <http://www.SAS%20institute\jmp4\help.>]
- Schroeder, Brian K. "Long-Range Forecast Possibilities for X-Band Radar Construction on Shemya," Master's Thesis, Air Force Institute of Technology, Department of Engineering Physics, 2-4, 2002.
- Ting, Mingfang, H. Wang, 1997: Summertime U.S. precipitation variability and its relation to Pacific sea surface temperature, *Journal of Climate*, **10**, 1853-1873.

Trenberth, K.E., 1981: *Characteristic Patterns of Variability of Sea Level Pressure in the Northern Hemisphere*, Monthly Weather Review, **109**, 1169-89.

Trenberth, K.E., 1991: General Characteristics, *Teleconnections Linking Worldwide Climate Anomalies*, M.H. Glantz, R.W. Katz, & N. Nicholls, Ed., Cambridge University Press, New York, 1-3.

U.S. CPC, cited 2002: Teleconnection Introduction. [Available online at <http://www.cpc.ncep.noaa.gov/data/teledoc/telecontents.html>.]

U.S.G.S. cited 2002: Learning Web. [Available online at [http://interactive2.usgs.gov/learning\\_web/textonly/teachers/globalchange.htm](http://interactive2.usgs.gov/learning_web/textonly/teachers/globalchange.htm).]

Wilks, D.S. 1995: *Statistical Methods in the Atmospheric Sciences*, Academic Press, 467pp.

## Vita

Richard is currently attending the Graduate Meteorology program, Department of Engineering Physics, Air Force Institute of Technology at Wright-Patterson Air Force Base, Ohio. Upon graduation, he will be assigned to an undisclosed DoD facility.

<b>REPORT DOCUMENTATION PAGE</b>			Form Approved OMB No. 074-0188		
<p>The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of the collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.</p> <p><b>PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.</b></p>					
<b>1. REPORT DATE (DD-MM-YYYY)</b> 12-02-2003		<b>2. REPORT TYPE</b> Master's Thesis		<b>3. DATES COVERED (From - To)</b> June 2002 - March 2003	
<b>4. TITLE AND SUBTITLE</b>  DATA MINING ATMOSPHERIC/OCEANIC PARAMETERS IN THE DESIGN OF A LONG-RANGE NEPHELOMETRIC FORECAST TOOL			<b>5a. CONTRACT NUMBER</b>		
			<b>5b. GRANT NUMBER</b>		
			<b>5c. PROGRAM ELEMENT NUMBER</b>		
<b>6. AUTHOR(S)</b>  Benz, Richard F., Major, USAF			<b>5d. PROJECT NUMBER</b>		
			<b>5e. TASK NUMBER</b>		
			<b>5f. WORK UNIT NUMBER</b>		
<b>7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)</b> Air Force Institute of Technology Graduate School of Engineering and Management (AFIT/ENP) 2950 Hobson Way, Building 640 WPAFB OH 45433-7765			<b>8. PERFORMING ORGANIZATION REPORT NUMBER</b>  AFIT/GM/ENP/03-02		
<b>9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)</b> 28th OWS 905 Patrol Rd. Shaw AFB SC 29152-5307 Major Walter Otto walter.otto@shaw.af.mil DSN 965-0500			<b>10. SPONSOR/MONITOR'S ACRONYM(S)</b>		
			<b>11. SPONSOR/MONITOR'S REPORT NUMBER(S)</b>		
<b>12. DISTRIBUTION/AVAILABILITY STATEMENT</b> APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.					
<b>13. SUPPLEMENTARY NOTES</b>					
<b>14. ABSTRACT</b> The Department of Defense calls for long-range forecasts to aid in the planning of operations. The goal of this research was to explore the feasibility of predicting, one month in advance, the total monthly cloud cover over the country of Afghanistan. Data was examined using standard statistical regression techniques, including linear and multiple linear regression, and then CART analysis was used for exploring possible concealed structure (due to current events, CART analysis was also applied to the country of Iraq; see Appendix C). Standard statistics showed a strong negative correlation between monthly average OLR and surface observational total cloud cover from the fall through spring months. However, linear regression revealed very weak relationships between the predictor and predictand variables. As well, CART results contained misclassification rates that exceeded established thresholds for operational use. Further studies using CART for atmospheric sciences applications should be pursued.					
<b>15. SUBJECT TERMS</b> Teleconnection indices (TIs), data mining, Classification and Regression Trees (CART), Afghanistan, total cloud cover, sea surface temperature (SST), outgoing longwave radiation (OLR), Real Time Nephanalysis (RTNEPH)					
<b>16. SECURITY CLASSIFICATION OF:</b>			<b>17. LIMITATION OF ABSTRACT</b>  UU	<b>18. NUMBER OF PAGES</b>  100	<b>19a. NAME OF RESPONSIBLE PERSON</b> Ronald, P. Lowther, Lt Col, USAF (ENP)
a. REPO RT	b. ABSTRACT	c. THIS PAGE			<b>19b. TELEPHONE NUMBER (Include area code)</b> (937) 255-3636, ext 4645
U	U	U			

Standard Form 298 (Rev. 8-98)  
Prescribed by ANSI Std. Z39-18