

Air Force Institute of Technology

**AFIT Scholar**

---

Theses and Dissertations

Student Graduate Works

---

3-2002

## Exploration of Teleconnection Indices for Long-Range Seasonal Temperature Forecasts

Robb M. Randall

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



Part of the [Meteorology Commons](#)

---

### Recommended Citation

Randall, Robb M., "Exploration of Teleconnection Indices for Long-Range Seasonal Temperature Forecasts" (2002). *Theses and Dissertations*. 4499.  
<https://scholar.afit.edu/etd/4499>

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 [AFIT.ENWL.Repository@us.af.mil](mailto:AFIT.ENWL.Repository@us.af.mil).



**EXPLORATION OF TELECONNECTION INDICES FOR LONG-RANGE  
SEASONAL TEMPERATURE FORECASTS**

THESIS

Robb M. Randall, Captain, USAF  
AFIT/GM/ENP/02M-08

**DEPARTMENT OF THE AIR FORCE  
AIR UNIVERSITY**

**AIR FORCE INSTITUTE OF TECHNOLOGY**

---

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

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

## Report Documentation Page

<b>Report Date</b> 11 Mar 02	<b>Report Type</b> Final	<b>Dates Covered (from... to)</b> Jun 01 - Mar 02
<b>Title and Subtitle</b> Exploration of Teleconnection Indices for Long-Range Seasonal Temperature Forecasts	<b>Contract Number</b>	
	<b>Grant Number</b>	
	<b>Program Element Number</b>	
<b>Author(s)</b> Captain Robb M. Randall, USAF	<b>Project Number</b>	
	<b>Task Number</b>	
	<b>Work Unit Number</b>	
<b>Performing Organization Name(s) and Address(es)</b> Air Force Institute of Technology Graduate School of Engineering and Management (AFIT/EN) 2950 P Street, Bldg 640 WPAFB OH 45433-7765	<b>Performing Organization Report Number</b> AFIT/GM/ENP/02M-08	
<b>Sponsoring/Monitoring Agency Name(s) and Address(es)</b> AFCCC/DOO ATTN: Mr. Ken Walters 151 Patton Ave. Rm 120 Asheville, NC 28801-5002	<b>Sponsor/Monitor's Acronym(s)</b>	
	<b>Sponsor/Monitor's Report Number(s)</b>	
<b>Distribution/Availability Statement</b> Approved for public release, distribution unlimited		
<b>Supplementary Notes</b> The original document contains color images.		

**Abstract**

The Air Force Combat Climatology Center (AFCCC) is tasked to provide long-range seasonal forecasts for worldwide locations. Currently, the best long-range temperature forecasts the weather community has are the climatological standard normals. This study creates a stepping-stone into the solution of long-range forecasting by finding a process to predict temperatures better than those using climatological standard normals or simple frequency distributions of occurrences. Northern Hemispheric teleconnection indices and the standardized Southern Oscillation index are statistically compared to three-month summed Heating Degree Days (HDDs) and Cooling Degree Days (CDDs) at 14 U.S. locations. First, linear regression was accomplished. The results showed numerous valid models, however, the percent of variance resolved by the models was rarely over 30%. The HDDs and CDDs were then analyzed with Data-mining classification tree statistics, however, the results proved difficult to extract any predictive quantitative information. Finally a Data-mining regression tree analysis was performed. At each conditional outcome, a range of HDDs/CDDs is produced using the predicted standard deviations about the mean. Verification of independent teleconnection indices was used as predictors in the conditional model; 90% of the resulting HDDs/CDDs fell into the calculated range. An overall average reduction in the forecast range was 35.7% over climatology

**Subject Terms**

Climatology, teleconnection indices, data mining, Classification and Regression Trees (ART), Heating Degree Days (HDD), Cooling Degree Days (CDD)

**Report Classification**

unclassified

**Classification of this page**

unclassified

**Classification of Abstract**

unclassified

**Limitation of Abstract**

UU

**Number of Pages**

87

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 U.S. Government.

AFIT/GM/ENP/02M-08

EXPLORATION OF TELECONNECTION INDICES FOR LONG-RANGE  
SEASONAL TEMPERATURE FORECASTS

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 of the Requirements for the

Degree of Master of Science in Meteorology

Robb M. Randall, B.S.

Captain, USAF

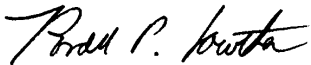
March 2002

APPROVED FOR PUBLIC RELEASE; DISTRUBUTION UNLIMITED


EXPLORATION OF TELECONNECTION INDICES FOR LONG-RANGE  
SEASONAL TEMPERATURE FORECASTS

Robb M. Randall, B.S.  
Captain, USAF

Approved:

  
\_\_\_\_\_  
Ronald P. Lowther (Chairperson)

6 MAR 02  
date

  
\_\_\_\_\_  
Michael K. Walters (Member)

6 MAR 02  
date

  
\_\_\_\_\_  
Edward D. White (Member)

8 Mar 02  
date

## Acknowledgments

God has entrusted us with the care of His resources on earth, to use wisely. In doing so I pray this study will be the springboard for conserving some of the energy that we use to run our daily operations and missions. I thank and praise God for leading me in a direction He felt necessary and surrounding me with people with the knowledge to guide me through this study.

The most important person in my life, my wife, has been the greatest friend and supporter of the work needed to complete this project. In that sense this study is part her work, and words can't explain the love I feel for her and the gratitude I have for her part in this project. I thank my children for reminding me what a wonderful blessing it is to be a father.

I thank my faculty advisor, Lt Col Ron Lowther, who's great passion for climatology could not help but rub off on me. His faith in a project that he wasn't sure could be solved, his knowledge, sense of humor and, most of all, his patience, with a difficult student are attributes I need to praise.

Thank you to the rest of my committee, the *approachable* Lt Col Mike Walters and Maj Tony White. Lt Col Walters knowledge of PCA was an excellent resource and Maj White's levels of statistical hell were all worth it in the end. I thank them both for their patience and time.

My thanks to my good friend, SMSgt Rex Ellis (Retired), for his professional and spiritual guidance through the years, and my brother for keeping me motivated in new technologies and the scientific frontier.



I'm very grateful to the Air Force Combat Climatology Center and especially, Mr. Ken Walters, for their sponsorship in this project. The Air Force Technical Library staff was a great help with this research and I thank them for their dedication and letting us invade their territory.

I thank Maj (sel) Todd McNamara and 1Lt Lee Nelson for their friendship and support; Capt Dino Carter, Capt Brian Schroeder, and 1Lt Hugh Freestrom for the great research group, and to the rest of the class, everyone was a supporter in one-way or another. Even though the red tide wasn't in this work, I thank Jeff Sitler for the great support he gave me during this project.

## Table of Contents

	Page
Acknowledgments.....	iv
List of Figures.....	viii
List of Tables.....	x
Abstract.....	xi
I. Introduction.....	1
Background.....	1
Long-range Weather Forecasting.....	2
Scope of Research.....	4
Research Objectives.....	5
II. Literature Review.....	6
Rotated Principle Component Analysis (RPCA).....	6
Northern Hemispheric Teleconnection patterns.....	7
Southern Hemispheric Teleconnection pattern.....	21
Other research.....	22
III. Data Collection and Review.....	23
Northern Hemisphere Teleconnection Pattern Indices .....	23
Southern Oscillation Index.....	23
Heating Degree Days / Cooling Degree Days.....	24
Locations.....	24
IV. Linear Regression Analysis.....	26
Data manipulation for Regression Analysis.....	26
Linear Regression Analysis.....	28
V. Tree Based Statistical Models.....	31
Overview.....	31
Classification Tree Analysis.....	32
Regression Tree Analysis .....	34
Application of Regression Tree Analysis.....	36
Results vs Frequency Distribution.....	38

VI.	Conclusions and Recommendations.....	43
	Conclusions.....	43
	Recommendations.....	45
	Bibliography.....	47
	Appendix. Regression Trees.....	49

## List of Figures

Figure	Page
1. Phases of NAO pattern.....	9
2. Phases of EA pattern.....	10
3. Phases of EA-JET pattern.....	11
4. Phases of EA/WR pattern.....	12
5. Phases of SCAND pattern.....	13
6. The POL pattern.....	14
7. Positive phase of ASU pattern.....	15
8. Phases of WP pattern.....	16
9. Phases of EP pattern.....	17
10. Phases of NP pattern.....	18
11. Phases of PNA pattern.....	19
12. Phases of the TNH pattern.....	20
13. Phases of PT pattern.....	21
14. Fourteen U.S. locations from which HDDs and CDDs are calculated.....	25
15. Example of a classification tree.....	33
16. Example of a regression tree.....	36
17. Example of a pruned regression tree.....	42

## List of Figures

Figure	Page
18. Atlanta regression tree.....	50
19. Chicago regression tree.....	52
20. Cincinnati regression tree.....	53
21. Dallas-Fort Worth regression tree.....	56
22. DeMoines regression tree.....	57
23. Las Vegas regression tree.....	60
24. Memphis regression tree.....	62
25. Minneapolis regression tree.....	63
26. New York, LaGuardia regression tree.....	65
27. Philadelphia regression tree.....	66
28. Portland regression tree.....	68
29. Sacramento regression tree.....	69
30. Tucson regression tree.....	71
31. WPAFB regression tree.....	73

## List of Tables

Table.....	Page
1. Monthly periods used in summations of HDDs and CDDs to create seasons.....	27
2. ANOVA table output from linear regression.....	28
3. Coefficients table output from linear regression.....	29
4. Model summery table output from linear regression.....	29
5. P-value and adjusted R-squared from the ANOVA table for the 14 locations.....	30
6. Shapiro-Wilk goodness of fit test for normality.....	37
7. Percentage of CDDs that were in the predicted range after verification data were run through tree.....	38
8. Expected forecast range reduction.....	40
9. The total expected forecast range reduction over climatology for the 14 forecast locations.....	41

## **Abstract**

The Air Force Combat Climatology Center (AFCCC) is continually tasked to provide temperature and other long-range seasonal forecasts for locations at which Department of Defense (DoD) personnel are performing long-range exercises and real-world mission planning support. DoD needs long-range forecasts to estimate how much fuel is necessary to keep energy production, purchases and operations at the proper levels to accommodate all the energy needs on their installations and within their worldwide theaters of operation. Currently, the best long-range temperature forecasts the weather community has for worldwide locations use either climatological standard normals or simple frequency distributions of occurrences. This study creates a stepping-stone toward the solution of long-range temperature forecasting by finding a process to predict more accurate temperatures than those forecasts obtained using climatological standard normals or simple frequency distributions of occurrences. This same solution is also highly sought after by many non-DoD users as well.

Northern Hemispheric teleconnection indices, created by rotated principle component analysis (RPCA), and the standardized Southern Oscillation index are statistically compared to Heating Degree Days (HDDs) and Cooling Degree Days (CDDs) at 14 U.S. locations. HDDs and CDDs were summed over three-month periods to compute seasonal summations. Teleconnection indices found to be leading modes, using RPCA, in a particular month are compared to the HDD/CDD summations of the following three months in order to create predictive models.

First, linear regression is accomplished on the data. The results show numerous valid modes, however, the percent of HDD and CDD variance resolved by the modes is rarely over 30%. The HDDs and CDDs are then categorized and analyzed with a classification tree data-mining program, however, the results did not show any predictive quantitative information.

A regression tree data mining analysis is then performed on the uncategorized HDDs/CDDs, which shows excellent conditional predictive outcomes. At each conditional outcome, a range of HDDs/CDDs is produced using the predicted standard deviations about the mean. When teleconnection indices were used as predictors in the conditional model, 90% of the time the resulting HDDs/CDDs fell into the calculated range. Expected forecast range reductions over climatology are then calculated, and an overall average expected forecast range reduction of 35.7% over climatology was achieved.



# EXPLORATION OF TELECONNECTION INDICES FOR LONG-RANGE SEASONAL TEMPERATURE FORECASTS

## **I. Introduction**

### **Background**

The Air Force Combat Climatology Center (AFCCC) is continually tasked to provide temperature forecasts for locations at which Department of Defense (DoD) personnel are performing long-range exercises and real-world mission planning support. The importance of these forecasts comes down to the cost of moving equipment and supplies, aircraft fuel loads, humanitarian assistance packages, and other operational needs. Commanders require accurate temperature forecasts in order to plan equipment resources necessary to keep troops safe from the environmental elements.

Any necessary equipment or clothing can drastically change the logistical requirements of any mission, which is measured in costs and expediency. For example, a mission anywhere where the temperature falls below freezing requires extra clothing, heating equipment, heated facilities, additional aircraft maintenance equipment, deicing equipment, etc. A large mission with these requirements can add millions of dollars to the cost of the deployment. The Gulf War, for example, cost 61 billion dollars (Horan, 1997). Troops were required to take both hot and cold weather clothing items for the variety of weather conditions experienced in the region (USAF, 1991). If it were possible to give commanders better long-range temperature forecasts, they might have been able

to alleviate taking all or part of the cold-weather clothing, saving millions of dollars and vitally needed airlift requirements in the process. Accurate forecasts can also help smaller scale military teams as well. For example, special operation forces deployed in a country such as Afghanistan cannot afford to carry unnecessary equipment. Both cold mountain areas and hot desert areas dominate the Afghanistan terrain. Accurate long-range forecasts are vital to their mission success as well as the success of the massive airlift operations required to support the war effort.

Long-range temperature forecasting is not only important in mission planning, but also for planning fuel costs for energy consumption. The DoD, just like the general population, needs to forecast how much fuel is necessary to keep energy production and purchases at the proper levels to accommodate all the energy needs on their installations and in their worldwide theaters of operation. This can become very difficult, especially if there are significant temperature anomalies, such as periods of extreme hot or cold conditions. When there are significant temperature anomalies, there is usually not enough fuel to maintain the amount of energy being consumed. The better the long-range temperature forecasts are, the better the initial estimates of needed fuel reserves for energy use. In addition, it is hoped the improvement of long-range temperature forecasts may lead to improved long-range forecasts of other climatic elements.

### **Long-range weather forecasting**

Lorenz saw his initial weather patterns grow farther and farther apart in model simulations until all resemblance to each other had disappeared. He decided that long-range weather forecasting must be doomed (Gleick, 1987). Today, it is thought that numerical models are not valid after the 15-day point (Anthes, 1986). Clearly the

immediate future of long-range weather forecasting does not lie with the use of short-range numerical weather prediction models.

Baur (1951) suggested long-range weather forecasting could be possible using large-scale spatial circulation patterns, which he termed *Grosswetterlagen*. Since then, countless studies compared large-scale weather patterns with weather parameters around the world. Most, however, do not try to use the patterns as forecast tools. This research attempts to look at forecasting long-range temperatures by using techniques similar to the *Grosswetterlagen* method, using global teleconnection patterns (Wallace and Gutzler, 1981). This research attempts to take the concept Baur had and use today's technology to make forecasts once thought impossible.

Currently, the best long-range temperature forecasts the weather community has for worldwide locations are the climatological standard normals, which are averages of climatological data calculated for the following consecutive 30-year periods, established by international agreement: 1 January 1901 to 31 December 1930; 1 January 1931 to 31 December 1960; 1 January 1961 to 31 December 1990; etc. (Glickman, 2001). The U.S. Climate Prediction Center (CPC) calculates standard normals for U.S. stations at the end of each decade (CPC, 2001). However, temperature anomalies, which are the most important features in long-range mission and energy planning, are smoothed out or unseen over such 30-year averages.

This research focuses on finding a process to predict more accurate temperatures than those obtained by using climatological standard normals or simple frequency distributions of occurrences. This study investigates the relationships between temperature and known global teleconnection patterns. Finding a significant relationship

that affects DoD missions and energy consumption might possibly save DoD billions of dollars per year.

### **Scope of Research**

Any forecasting tool needs to be reproducible and readily available for users in the field, without a great deal of trouble gaining necessary data. For this reason, the National Center for Environmental Prediction's (NCEP) CPC's Standardized Northern Hemisphere Teleconnection Indices and the Southern Oscillation Index are used in this research. CPC's indices are produced monthly and are available to users on their web site: <http://www.cpc.ncep.noaa.gov>. This research investigates statistical methods of using these monthly indices to predict U.S. seasonal temperatures from one to three months in advance.

One way to represent temperature forecasts over a period of time is taken from the civil engineering community. Their primary need is a means to relate temperatures to the demand for fuel consumption over a specific period of time, and they utilize Heating Degree-Days (HDDs) or Cooling Degree-Days (CDDs) in this effort.

This research uses various statistical software packages to explore any relationships between teleconnection indices and HDDs/CDDs for 14 locations that have current temperature data available. To ensure the utmost quality of the temperature data used, only U.S. first-order stations are used in this analysis. All of the cities have different periods of record for their temperature data, and the teleconnection data is only from 1950 to present.

## **Research Objectives**

The goal of this research is to use known significant teleconnection indices to create a predictive tool for forecasting long-range temperature patterns over the U.S. The specific objectives necessary to achieve this goal are:

1. to gather temperature data from 14 locations across the U.S. in order to represent most climatic regimes across the country;
2. to calculate and compile monthly HDDs and CDDs values from this data;
3. to gather teleconnection indices from the 14 most significantly known Northern Hemisphere teleconnections and the Southern Oscillation Index in the Southern Hemisphere;
4. to remove ten years of the data for later verification of any relationship identified;
5. to analyze data with a thorough regression analysis to find any significant relationships between monthly teleconnection indices and the summation of HDDs/CDDs for the following three months;
6. to use, if necessary, data mining techniques to find any predictive relationships if standard statistical methods fail;
7. to create predictive tools using monthly teleconnection indices as the predictor and summed HDDs and CDDs seasons as the predictand for any relationships found;
8. to verify any predictive models developed by using ten years of independent data not included in creating the predictive models and,
9. to investigate the spatial homogeneity of the created prediction trees.

## **II. Literature Review**

### **Rotated Principle Component Analysis (RPCA)**

The method used for defining the low-frequency teleconnection patterns in this study is that of Rotated Principal Component Analysis (RPCA). RPCA is considered to be superior to using distinct centers of geopotential height anomalies at select locations, in that the teleconnection patterns identified are based on the entire flow field, and not just from height anomalies at the selected locations (Rodionov and Assel, 2000).

RPCA uses the eigenvectors of the cross-correlation (or cross-covariance) matrix from the time variations of the grid-point values of the 700-mb height anomalies, and ranks the eigenvectors according to the amount of total variance they explain (creating a PCA). The PCA is then orthogonally rotated to get the variances as close to zero as possible (Barnston and Livezey, 1987). Barnston and Livezey (1987) used the RPCA technique to calculate the 10 most prominent teleconnection patterns in each month. This procedure isolates the primary teleconnection patterns for all months and allows for a time series of the amplitudes of the patterns to be constructed.

CPC uses the Barnston and Livezey method by applying the RPCA technique to monthly mean 700-mb height anomalies between January 1964 and July 1994. In CPC's analysis, ten patterns are determined for each calendar month by using all of the height anomaly fields for the three-month period centered on that month. For example, the July patterns are calculated based on the June through August anomaly fields (CPC, 2001). Using RPCA instead of PCA creates solutions that have a physical meteorological interpretability. The RPCA solutions also involve much smaller areas of the hemisphere

(Barnston and Livezey, 1987). A more comprehensive discussion of rotated principal component solutions is found in Horel (1981) and Barnston and Livezey (1987).

### **Northern Hemispheric Teleconnection patterns**

Teleconnection patterns are macro- $\beta$  scale patterns resembling standing waves with geographically fixed centers (Horel, 1981). They are also referred to as preferred modes of low-frequency variability (CPC, 2001), and several teleconnection patterns in planetary circulation have been documented by Barnston and Livezey (1987). A comprehensive re-analysis of Northern Hemispheric variability patterns has been undertaken by CPC using newly available 700hPa height data (Washington et al., 2000) to achieve a better understanding in the synoptic weather patterns related to the teleconnection patterns.

The 13 prominent Northern Hemispheric teleconnection patterns used in this study are separated into three regions; patterns over the North Atlantic, patterns over Eurasia, and patterns over North Pacific/ North America. The prominent patterns over the North Atlantic are: the North Atlantic Oscillation (NAO), the East Atlantic Pattern (EA), and the East Atlantic Jet Pattern (EA-JET). The prominent patterns over Eurasia are: the East Atlantic/West Russia Pattern (EA/WR), the Scandinavian Pattern (SCAD), the Polar/Eurasia Pattern (POL) and the Asian Summer Pattern (ASU). The prominent patterns over the North Pacific/North America are: West Pacific Pattern (WP), the East Pacific Pattern (EP), the North Pacific Pattern (NP), the Pacific/North American Pattern (PNA), the Tropical/Northern Hemisphere Pattern (TNH), and the Pacific Transition Pattern (PT).

The North Atlantic Oscillation (NAO), shown in Figure 1, is one of the dominant modes of Northern Hemispheric climate variability (Walker and Bliss, 1932; Van Loon and Rogers, 1978; Wallace and Gutzler, 1981; Washington et al., 2000) and is a leading mode in all months (Barnston and Livezey, 1987; Washington et al., 2000). The NAO exhibits little variation in its climatological mean structure from month-to-month, and consists of a north-south dipole of anomalies, with one center over the Greenland/Iceland region and the other center, of opposite sign, spanning the central latitudes of the North Atlantic around the Azores between 35°N and 40°N. The positive phase of the NAO reflects below-normal heights and pressure across the high latitudes of the North Atlantic and above-normal heights and pressure over the central North Atlantic, the eastern United States and Western Europe. The negative phase reflects an opposite dipole pattern of height and pressure anomalies over these regions (Washington et al., 2000; CPC, 2001). Strong positive phases of the NAO tend to be associated with above-normal temperatures in the eastern United States and across northern Europe and with below-normal temperatures in Greenland and oftentimes across southern Europe and the Middle East (CPC, 2001).

The EA pattern, shown in Figure 2, is a prominent mode of low-frequency variability over the North Atlantic. It is a prominent mode in all months except May-August. It consists of a north-south dipole of anomaly centers, which span the entire North Atlantic Ocean from east to west with the zero line always positioned over England or France. The EA pattern is structurally similar to the NAO pattern; however, the anomaly centers are displaced southeastward to the approximate nodal lines of the NAO



pattern. The lower-latitude center contains a strong subtropical link, reflecting large-scale modulation in the strength and location of the subtropical ridge (CPC, 2001).

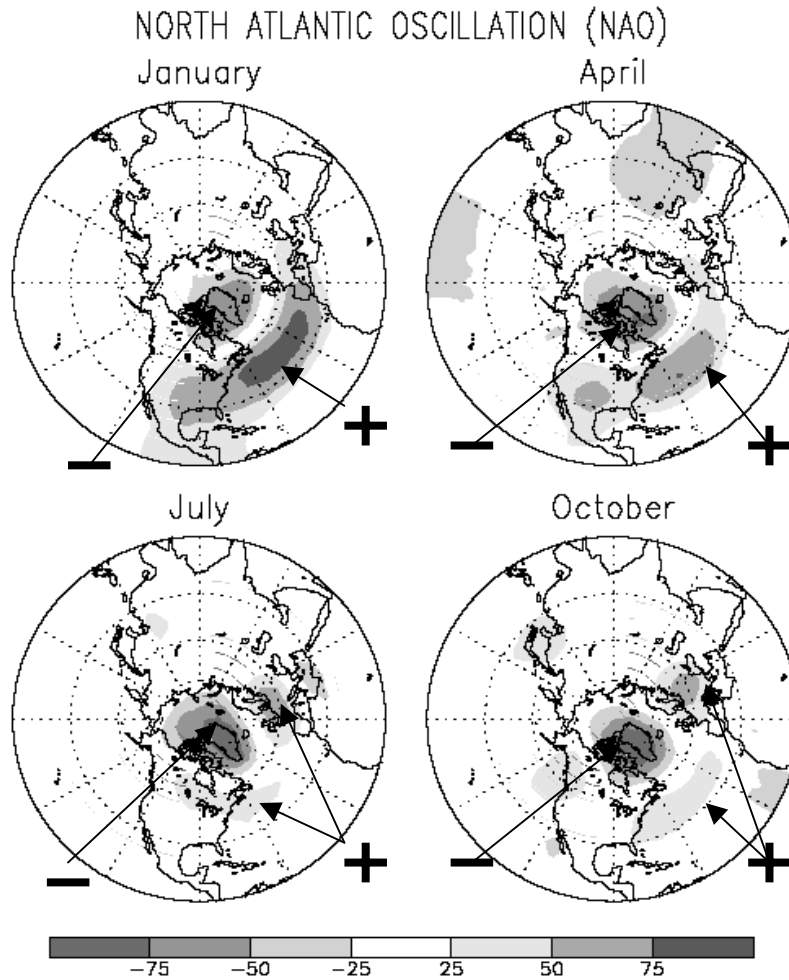


Figure 1. Phases of the NAO pattern. From positive phase in January to negative phase in July. Values are scaled to be correlations between the average 700-mb height anomalies at a given grid point and the principal component amplitude (modified from CPC, 2001).

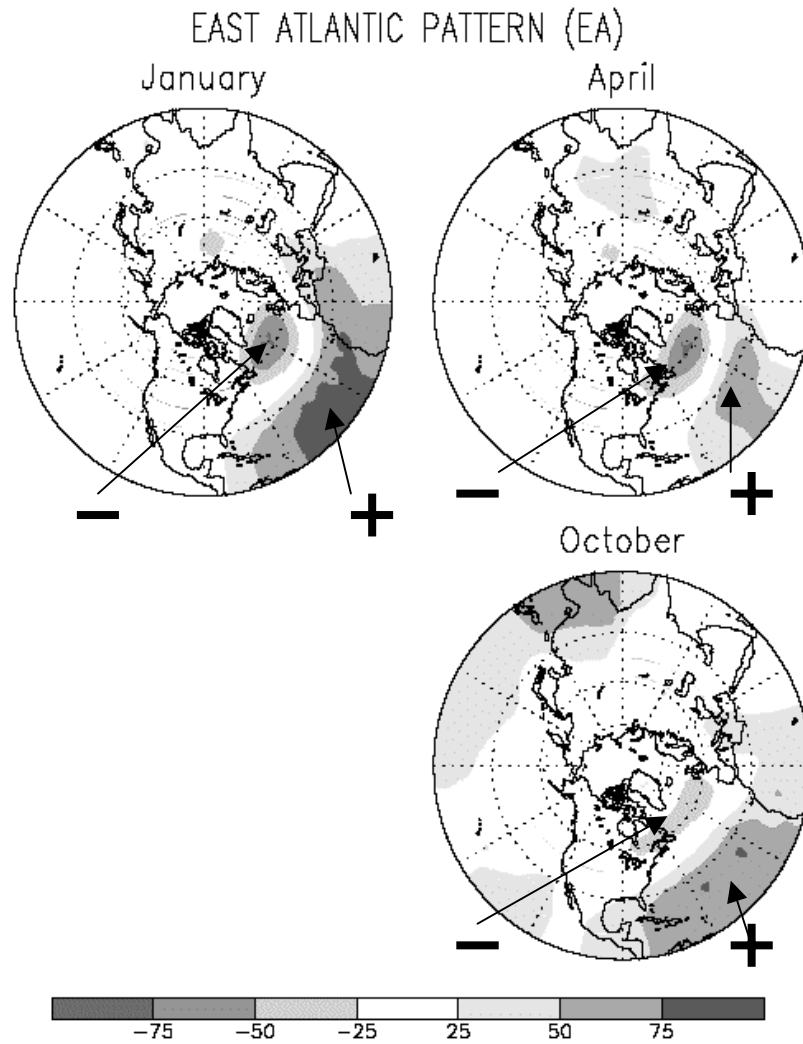


Figure 2. Phases of the EA pattern (modified from CPC, 2001).

The EA-JET pattern, shown in Figure 3, is a prominent mode of North Atlantic variability, appearing between April and August. This pattern also consists of a north-south dipole of anomaly centers, with one main center located over the high latitudes of the eastern North Atlantic and Scandinavia, and the other center located over Northern Africa and the Mediterranean Sea. A positive phase of the EA-Jet pattern reflects an intensification of the westerlies over the central latitudes of the eastern North Atlantic

and over much of Europe, while a negative phase reflects a strong split-flow configuration over these regions, sometimes, in association with long-lived blocking anticyclones in the vicinity of Greenland and Great Britain (CPC, 2001).

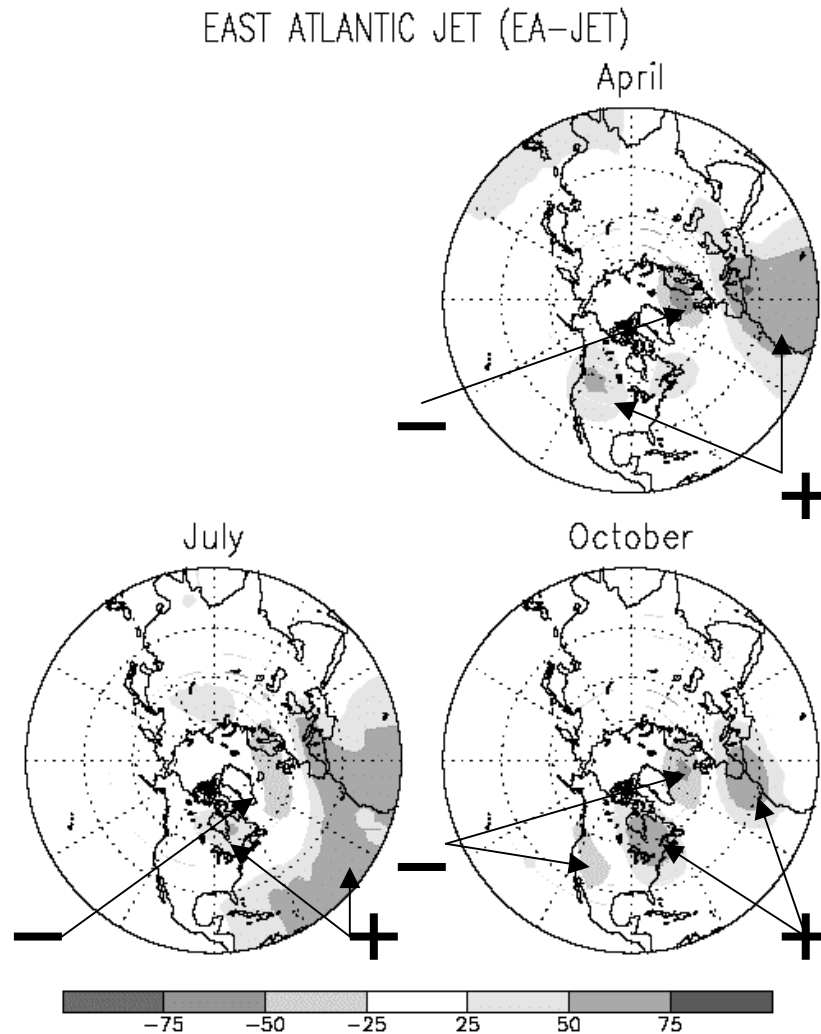


Figure 3. Phases of the EA-JET pattern (modified from CPC, 2001).

The EA/WR pattern, shown in Figure 4, is one of two prominent modes that affect Eurasia during most of the year. This pattern is prominent in all months except June-August. In winter, two main anomaly centers, located over the Caspian Sea and

Western Europe, comprise the East Atlantic/West Russian pattern. A three-celled pattern is then evident in the spring and fall seasons, with two main anomaly centers of opposite sign located over western/northwestern Russia and over northwestern Europe. The third center, having the same sign as the Russia center, is located off the Portuguese coast in spring, but exhibits a northern movement toward Newfoundland in the fall (CPC, 2001).

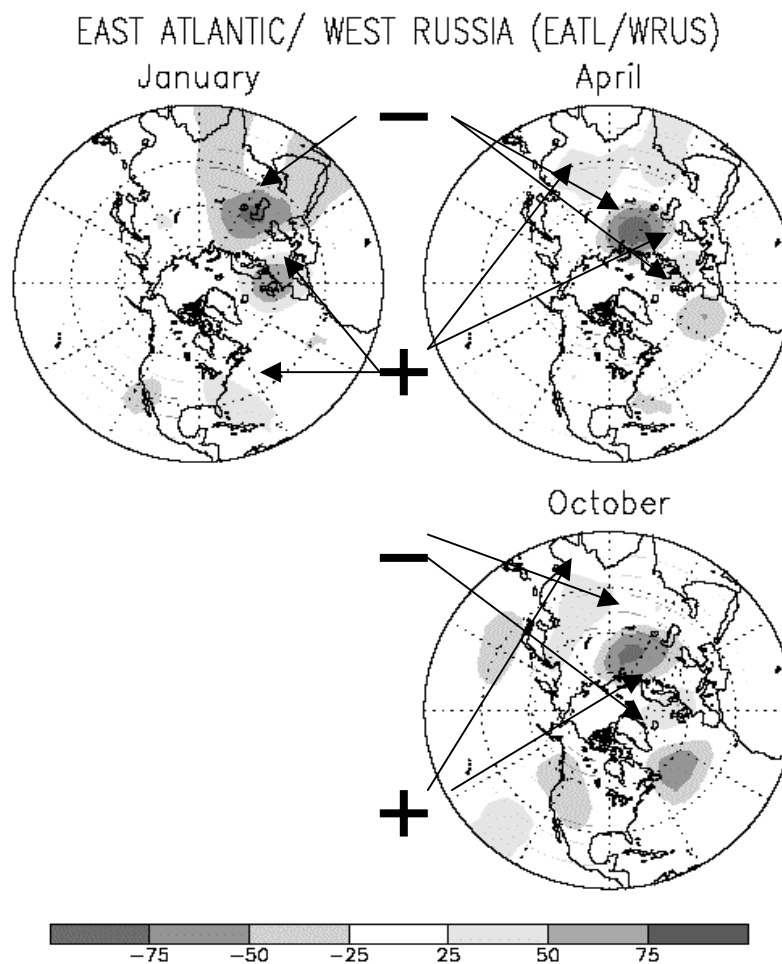


Figure 4. Phases of the EA/WR pattern (modified from CPC, 2001).

The SCAND pattern, shown in Figure 5, consists of a primary circulation center, which spans Scandinavia and large portions of the Arctic Ocean north of Siberia. Two additional weaker centers with opposite sign to the Scandinavia center are located over Western Europe and over the Mongolia and the western China sector. The positive phase of this pattern is associated with positive height anomalies, sometimes reflecting major blocking anticyclones over Scandinavia and western Russia, while the negative phase of the pattern is associated with negative height anomalies over these same regions (CPC, 2001).

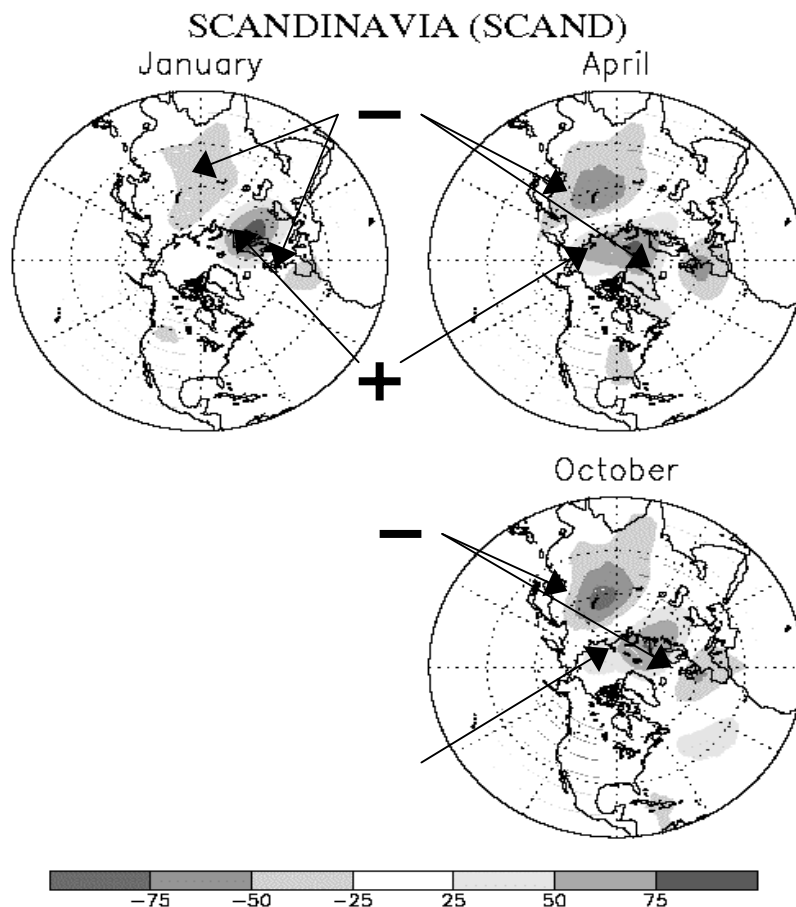


Figure 5. Phases of SCAND pattern (modified from CPC, 2001).

The POL pattern, shown in Figure 6, appears only in the winter, and is the most prominent mode of low-frequency variability during December and February. The pattern consists of one main anomaly center over the polar region, and separate centers of opposite sign to the polar anomaly over Europe and northeastern China. Thus, the pattern reflects major changes in the strength of the circumpolar circulation, and reveals the accompanying systematic changes that occur in the midlatitude circulation over large portions of Europe and Asia (CPC, 2001).

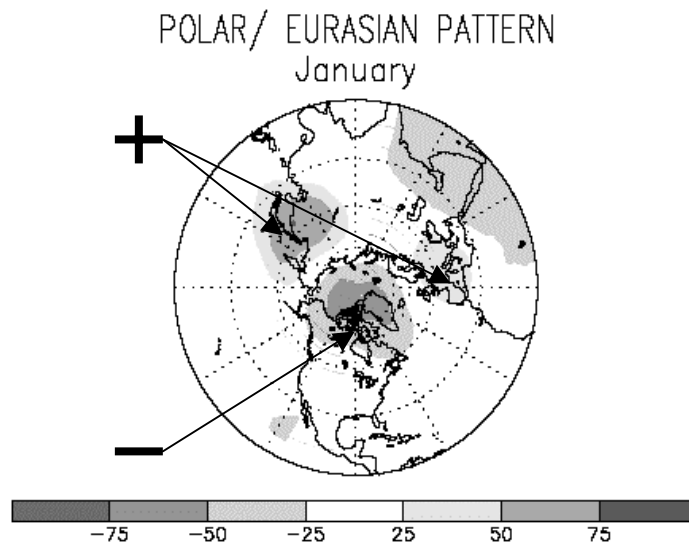


Figure 6. The POL pattern (modified from CPC, 2001).

The ASU pattern, shown in Figure 7, is a broad, east-west center in central Asia (Barnston and Livezey, 1987). The Asian Summer pattern is only a leading mode during the summer months of June-August. The pattern is monopole in nature with anomalies of the same sign observed throughout southern Asia and northeastern Africa. A positive

phase of the pattern is indicated by above-normal heights throughout southern Asia and northeastern Africa (CPC, 2001). The above normal heights are thought to be due to the intense heating over the Tibetan Plateau. It is theorized that in years with higher amounts of insolation over the plateau, the entire ITCZ over Africa and Asia is pulled further north thus affecting the circulation over the entire Asian continent (Lowther, 1998).

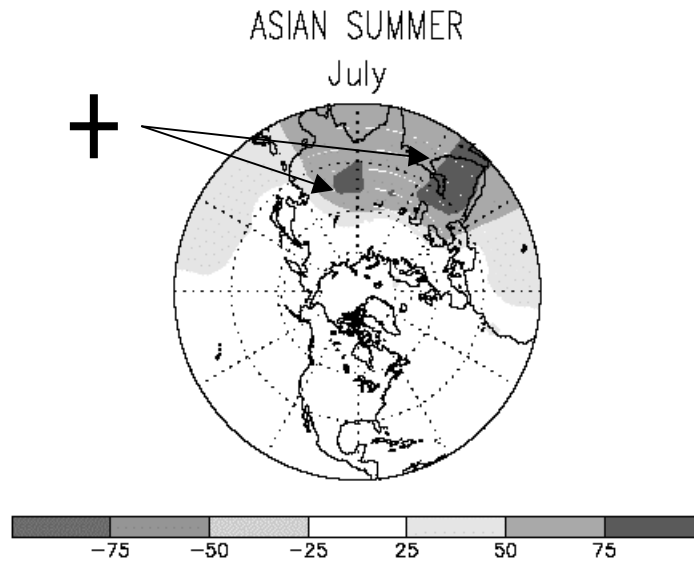


Figure 7. Positive phase of ASU pattern (modified from CPC, 2001).

The WP pattern, shown in Figure 8, is a primary mode of low-frequency variability over the North Pacific throughout all months (Washington et al., 2000; Barnston and Livezey, 1987; Wallace and Gutzler, 1981). During winter and spring, the pattern consists of a north-south dipole of anomalies, with one center located over the Kamchatka Peninsula and another broad center of opposite sign covering portions of southeastern Asia and the lower latitudes of the extreme western North Pacific. Strong positive or negative phases of this pattern reflect pronounced zonal and meridional

variations in the location and intensity of the entrance region of the Pacific (or East Asian) jet stream (CPC, 2001).

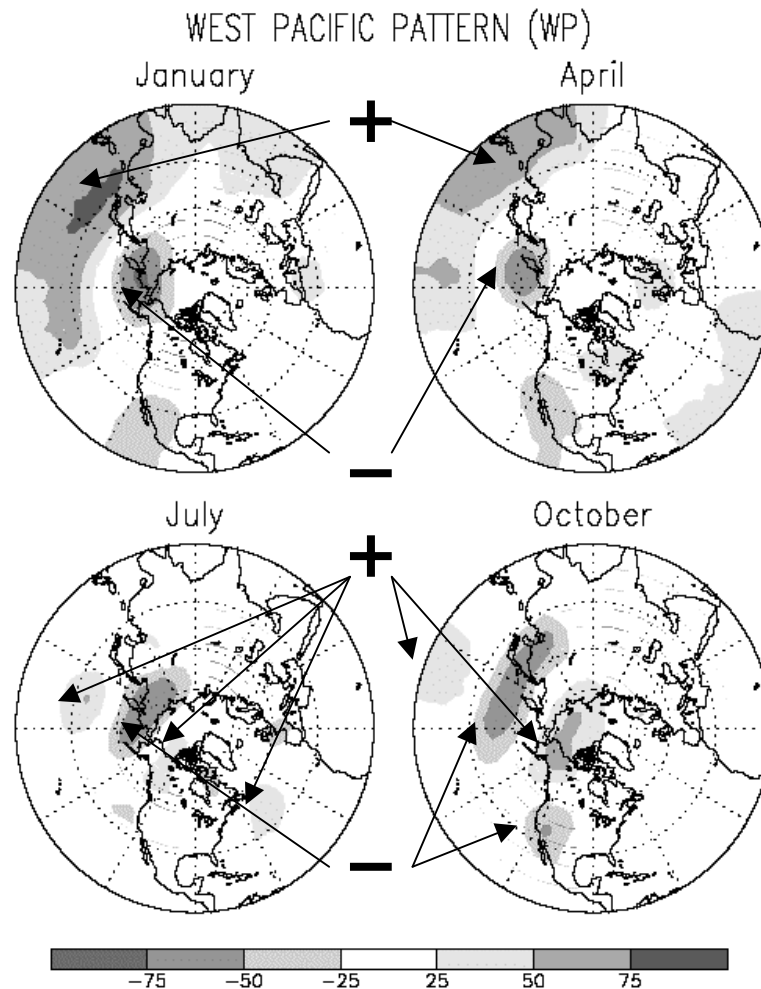


Figure 8. Phases of the WP pattern (modified from CPC, 2001).

The EP pattern, shown in Figure 9, is evident in all months except August and September and reflects a north-south dipole of height anomalies over the eastern North Pacific. The northern center is located in the vicinity of Alaska and the west coast of Canada, while the southern center is of an opposite sign and is found near, or east of,



Hawaii. During strong positive phases of the EP pattern, a deeper than normal trough is located in the vicinity of the Gulf of Alaska or western North America, and positive height anomalies are observed further south. A strong negative phase of the EP pattern is associated with a pronounced split-flow configuration over the eastern North Pacific, with reduced westerlies over the region (CPC, 2001).

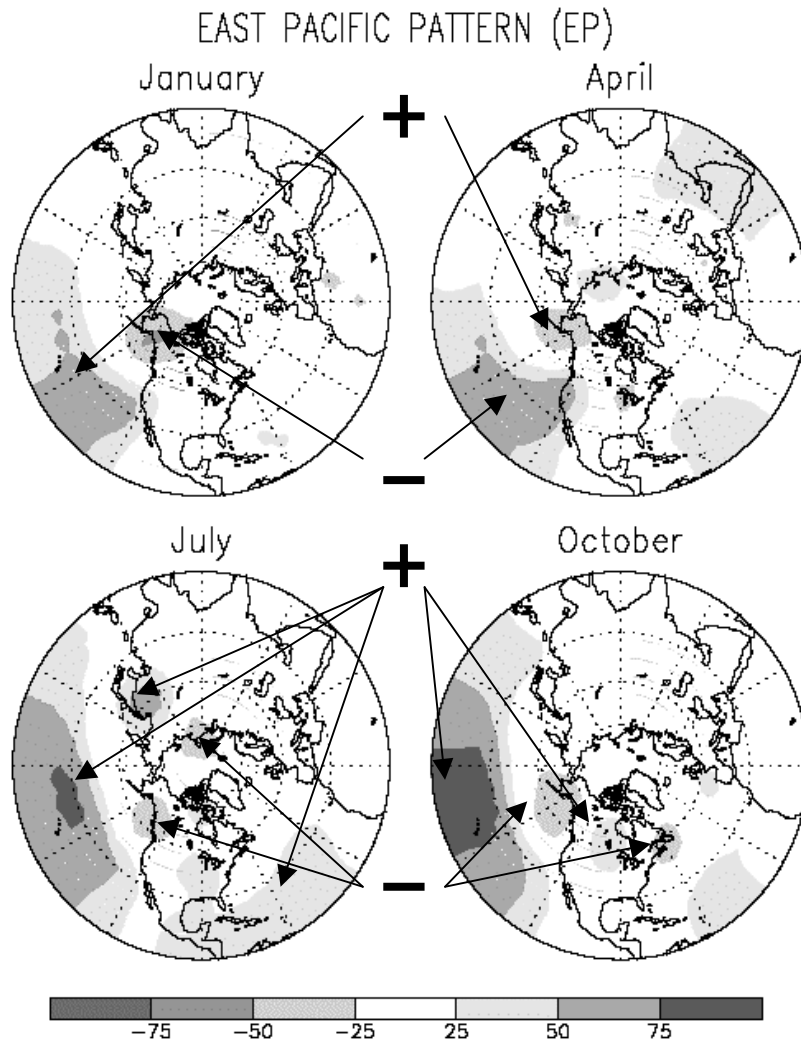


Figure 9. Phase of the EP pattern (modified from CPC, 2001).

The NP pattern, shown in Figure 10, is prominent from March through July. This pattern consists of a primary anomaly center, which spans the central latitudes of the western and central North Pacific, and a weaker anomaly region of opposite sign, which spans eastern Siberia, Alaska, and the western mountain regions of North America. Overall, pronounced positive phases of the NP pattern are associated with a southward shift and intensification of the Pacific jet stream from eastern Asia to the eastern North Pacific, followed downstream by an enhanced anticyclonic circulation over western North America, and by an enhanced cyclonic circulation over the southeastern United States. Pronounced negative phases of the NP pattern are associated with circulation anomalies of opposite sign in these same regions (CPC, 2001).

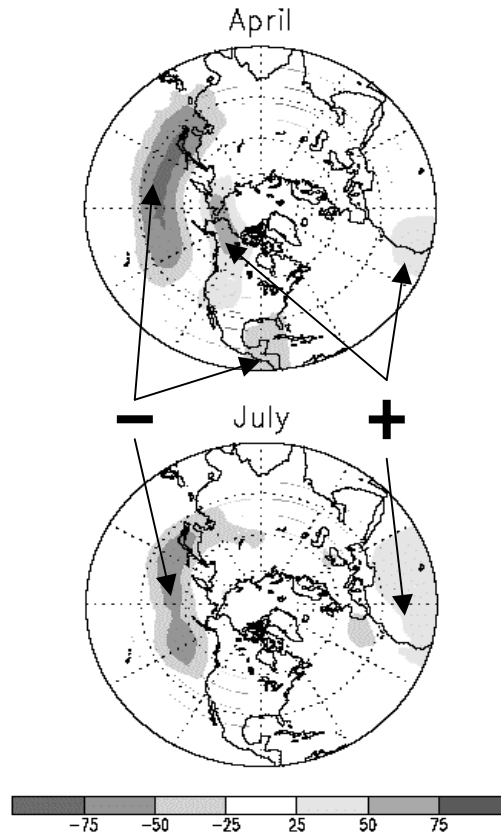


Figure 10. Phases of NP pattern (modified from CPC, 2001).

The PNA pattern, shown in Figure 11, is perhaps the best-known mode of Pacific-based variability. It appears in all months except June and July. The PNA pattern reflects a quadrupole pattern of geopotential anomalies, with anomalies of similar sign located south of the Aleutian Islands and over the southeastern USA. Anomalies with signs opposite to the Aleutian center are located near Hawaii and over central Canada during the winter and autumn (CPC, 2001; Washington et al., 2000; Barnston and Livezey, 1987).

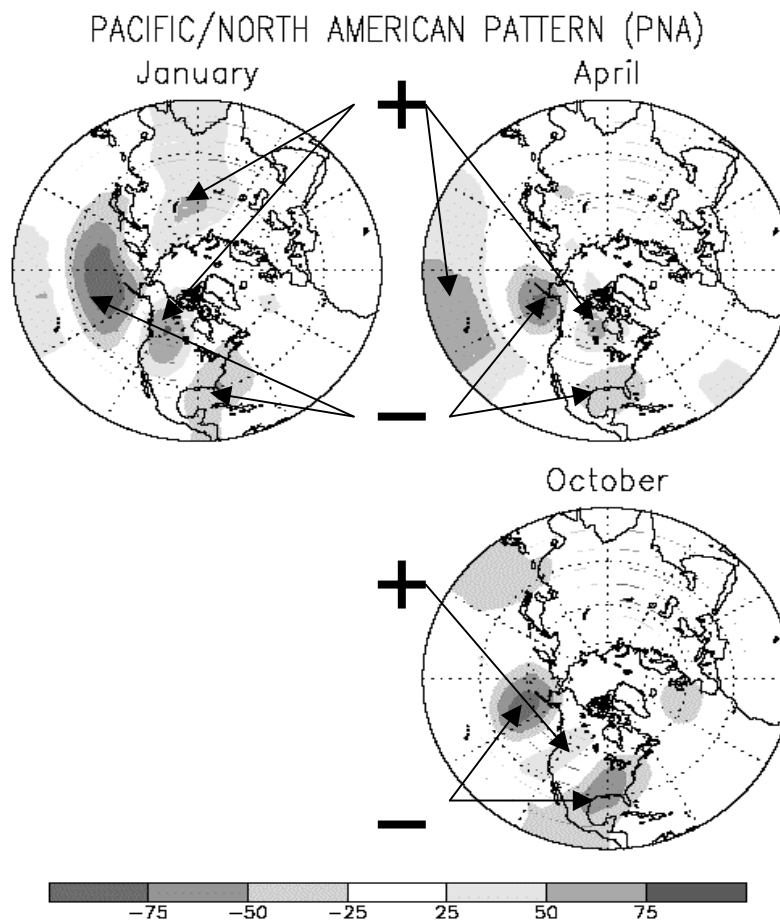


Figure 11. Phases of PNA pattern (modified from CPC, 2001).

The TNH pattern, shown in figure 12, appears as a prominent mode from November-February. The pattern consists of one primary anomaly center over the Gulf of Alaska and a separate anomaly center of opposite sign over Hudson Bay. A weaker area of anomalies having the same sign to the Gulf of Alaska anomaly extends across Mexico and the extreme southeastern United States. This pattern reflects large-scale changes in both the location and eastward extent of the Pacific jet stream, and also in the mean strength and position of the climatological Hudson Bay low. This pattern significantly modulates the flow of marine air into North America, as well as the southward transport of cold Canadian air into the north-central U. S. (CPC, 2001).

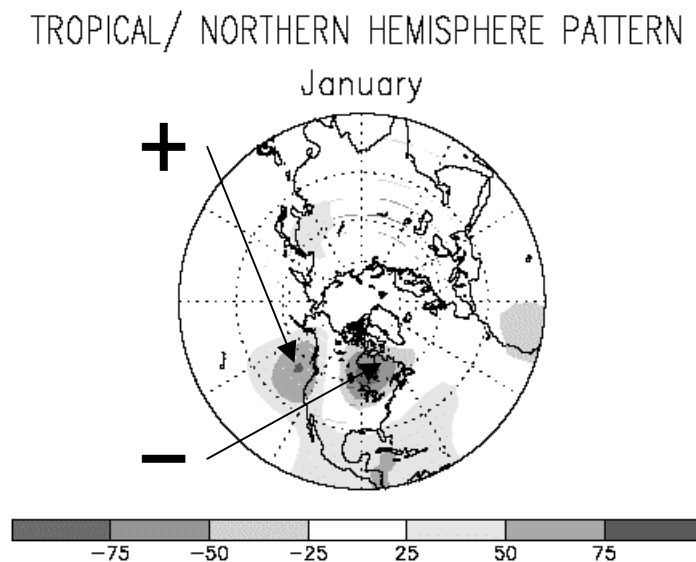


Figure 12. Phases of TNH pattern (modified by CPC, 2001).

The PT pattern, shown in Figure 13, is prominent between May-August. The mode consists of a pattern of height anomalies, which extends from the Gulf of Alaska

eastward to the Labrador Sea and is aligned along the 40°N latitude circle. The prominent centers of action have a similar sign and are located over the intermountain region of the United States and over the Labrador Sea. Relatively weak anomaly centers with signs opposite to the above are located over the Gulf of Alaska and over the eastern United States (CPC, 2001).

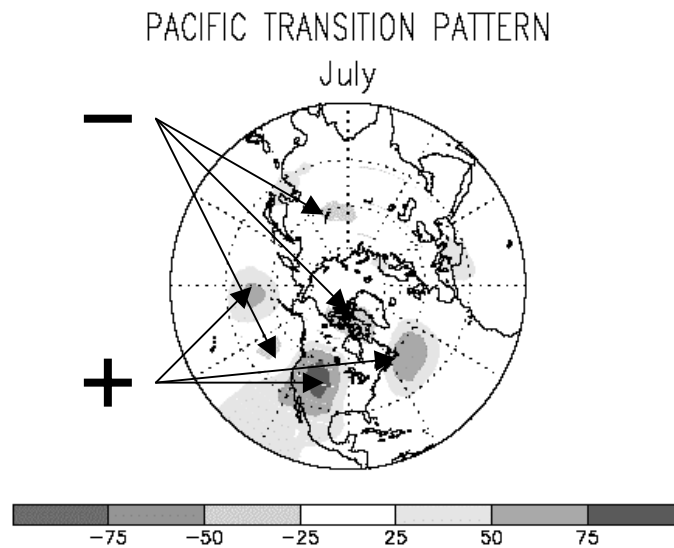


Figure 13. Phases of PT pattern (modified by CPC, 2001).

### **Southern Hemispheric Teleconnection Pattern**

“When the pressure is high in the Pacific Ocean, it tends to be low in the Indian Ocean from Africa to Australia.” This is how Sir Gilbert Walker described the Southern Oscillation (SO) in his papers in the 1920s and 1930s (Burroughs, 1992). There are numerous ways of recording this slow see-saw of atmospheric pressure across the equatorial pacific resulting in various Southern Oscillation Indexes (SOI). This research uses the SOI index created by the pressure difference between Tahiti, French Polynesia in

the mid-Pacific and Darwin in northern Australia. These two stations represent the Southeast Pacific area of high pressure and the Indonesian low, respectively (Robinson and Henderson-Sellers, 1999).

### **Other research**

There are numerous articles that draw comparisons between a specific teleconnection pattern and specific meteorological parameters, but there are fewer articles that use all of the teleconnections together for a comparison toward single parameters. Of those that use multiple teleconnections (Washington et al., (2000), Rodionov and Assel, (2000), for example), none attempted to create predictive relationships between the teleconnections and the parameters, thus resulting in a model to use as a tool in the operational field.

### III. Data Collection and Review

#### Northern Hemisphere Teleconnection Pattern Indices

As mentioned in Chapter II, the 10 most prominent teleconnection patterns in each month were calculated by RPCA (CPC, 2001). For each of the 10 patterns in a month, CPC calculates a monthly index. This method of calculation is a form of factor analysis that has not yet been published by CPC.

#### Southern Oscillation Index

The Southern Oscillation Index (SOI) is the only index used in this study that is not calculated by the RPCA method. It is calculated by using the raw atmospheric pressure data from Tahiti and Darwin, Australia. The anomalies used are departures from the 1951-1980 base period, and the anomaly for each city is defined as:

$$XA = (Actual(SLP)) - (mean(SLP)) \quad (1)$$

where the XA is either TA for the Tahiti anomaly or DA for the Darwin anomaly, depending on which cities anomaly is being calculated, and SLP is for the appropriate location sea level pressure. The standard deviation for Tahiti or Darwin is:

$$\text{Standard Deviation} = \sqrt{\frac{\sum XA^2}{N}} \quad (2)$$

where N is the number of months being summed. The data are then standardized as follows:

$$(ST/SD) = \frac{XA}{\sqrt{\frac{\sum XA^2}{N}}} \quad (3)$$

where ST is standardized Tahiti and SD is standardized Darwin monthly data.

The standardized SOI, is then:

$$SOI = \frac{ST - SD}{\sqrt{\frac{\sum (ST - SD)^2}{N}}} \quad (4)$$

where the denominator is the monthly standard deviation.

### **Heating Degree Days / Cooling Degree Days**

To calculate the HDDs for a particular day, one would first find the day's average temperature. The day's average temperature for the data used in this study is found by:

$$\frac{(Max\_temp + Min\_temp)}{2}, \text{ where the Max\_temp is the day's maximum temperature and}$$

Min\_temp is the day's minimum temperature. If the average temperature is less than 65°F, subtract the average temperature from 65°F and the result is the number of HDDs for that particular day. The resulting number is accumulated over a month, season, or whatever period is being examined. To calculate CDDs for a particular day, one would again find the day's average temperature. If the temperature is greater than 65°F, subtract 65°F from the average temperature and the result is the number of CDDs for that day. The number is again accumulated over the period in question.

### **Locations**

The HDDs and CDDs were calculated for 14 locations across the U.S., shown in Figure 14. The locations have a good history of temperature data and make an excellent database for this study. The locations are: Atlanta-Hartsfield International Airport, Georgia; Chicago O'Hare International Airport, Illinois; Cincinnati-Northern Kentucky Airport, Kentucky; Dallas-Fort Worth International Airport, Texas; Des Moines International Airport, Iowa; Las Vegas McCarran International Airport, Nevada;



Memphis International Airport, Tennessee; Minneapolis-St. Paul International Airport, Minnesota; New York Laguardia Airport, New York; Philadelphia International Airport, Pennsylvania; Portland International Airport, Oregon; Sacramento Executive Airport, California; Tucson International Airport, Arizona; and Wright-Patterson AFB, Ohio.



Figure 14. Fourteen U.S. locations from which HDDs and CDDs are calculated (modified from Mapquest.com, 2001).

## **IV. Linear Regression Analysis**

### **Data Manipulation for Linear Regression Analysis**

This study began with simple regression analysis between the HDDs and CDDs for the 14 locations and all 13 teleconnection indices. The goal was to compare the 13 teleconnection indices with the HDDs and CDDs of the 14 locations for one, two and three months in the future. All HDDs, CDDs, and teleconnections were put into data vector columns, temporally from January 1950 to December 1999, for 13 of the 14 locations. Chicago's data started in 1959, therefore Chicago data manipulations were accomplished from this date forward. The vector format used was necessary for the statistical program to properly accomplish regression analysis, but created missing data problems. The monthly teleconnection indices are created only in months the teleconnections are an RPCA leading mode (in the top ten). Except for the NAO and SOI standardized (SOI\_S), none of the teleconnections are in a leading mode every month of the year, thus the statistical will not use the data if there is missing data in any row of the combined columns.

To correct these problems 12 different matrices were created, one for each month of the year, and only those teleconnection indices that were RPCA leading modes in the specific month were added to the matrix. Needing to compare all twelve months of teleconnections with each location's HDDs and CDDs for one, two, and three months in the future significantly, increased the needed analysis time. Therefore due to time constraints, the data were combined to create seasonal values.

The months were combined to create seasons and then the seasons were separated into categories, shown in Table 1.

Table 1. Monthly periods used in summations of HDDs and CDDs to create seasons.

<i>Winter (HDD's)</i>		<i>Summer (CDD's)</i>	
October-December	<b>OND</b>	April-June	<b>AMJ</b>
November-January	<b>NDJ</b>	May-July	<b>MJJ</b>
December- February	<b>DJF</b>	June-August	<b>JJA</b>
January-March	<b>JFM</b>	July-September	<b>JAS</b>
February-April	<b>FMA</b>	August-October	<b>ASO</b>
March-May	<b>MAM</b>	September-November	<b>SON</b>

HDDs were summed into three-month seasons from October-May and CDDs were summed into three-month seasons from April-November. This process decreased the number of needed comparisons. The goal, at this point, was to compare the teleconnection indices in RPCA leading modes in a particular month with the summation of the next three month's HDDs or CDDs, depending on the month being compared for each location. Before these comparisons were completed, 10 years of data were randomly selected and removed to create an independent verification database.

### **Linear Regression Analysis**

Linear regression was accomplished on the data using leading mode teleconnections from May and the summed CDDs from June-August. The first output

statistic taken into consideration was the value in the significance column in the analysis of variance (ANOVA) table, shown in Table 2. This is commonly known as the p-value in statistical references and can be compared to the significance level of 0.01. If the p-value is less than 0.01 then at least one of the predictors (teleconnection indices) creates a statistically good model for the dependent variable (HDDs or CDDs) at the 0.01 significance level.

Table 2. ANOVA table output from linear regression. A p-value in the *Sig* column of less than 0.01 indicates a good model.

ANOVA						
Model		Sum of Squares	df	Mean Square	F	Sig.
	Regression	679422.475	10	67942.247	3.825	<b>.002</b>
	Residual	515055.500	29	17760.534		
	Total	1194477.975	39			

a Predictors: (Constant), SOI\_S, PNA, EAWR, PT, NAO, SCA, EA\_JET, NP, EP, WP  
b Dependent Variable: NUM\_AT

The value in the significant column of the coefficients table, shown in Table 3, is used to evaluate which predictors were statistically sound. Those with p-value greater than 0.05 were eliminated from the model and linear regression was rerun. This procedure was repeated until the best model was gained. Ideally, an ANOVA p-value of less than or equal to 0.01 with p-values of the predictors in the coefficients table of less than or equal to 0.01 result in the best model; however, it was not always possible to reach this goal. While running the linear regression, the Adjusted R-squared parameter in the model summary table, shown in Table 4, was considered.

Table 3. Coefficients table output from linear regression. A p-value of 0.01 in the *Sig* column is desired for a significant model. The predictors with the greatest p-value were eliminated and the analysis was run again.

Coefficients						
		Unstandardized Coefficients		Standardized Coefficients	t	<i>Sig.</i>
Model		B	Std. Error	Beta		
1	(Constant)	1187.300	23.599		50.311	<b>.000</b>
	NAO	-12.627	23.333	-.069	-.541	<b>.593</b>
	EA_JET	-.907	25.697	-.005	-.035	<b>.972</b>
	WP	91.944	25.580	.531	3.594	<b>.001</b>
	EP	36.439	28.100	.174	1.297	<b>.205</b>
	NP	56.544	22.089	.367	2.560	<b>.016</b>
	PNA	-25.205	23.185	-.138	-1.087	<b>.286</b>
	EAWR	-67.043	25.475	-.335	-2.632	<b>.013</b>
	SCA	-22.762	20.866	-.139	-1.091	<b>.284</b>
	PT	-17.405	25.303	-.098	-.688	<b>.497</b>
	SOI_S	43.276	30.210	.221	1.432	<b>.163</b>

a. Dependent Variable: NUM\_AT

Table 4. Model summary table output from linear regression. An Adjusted R-squared greater than 0.60 is desired.

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.754 <sup>a</sup>	.569	.420	133.2687

a. Predictors: (Constant), SOI\_S, PNA, EAWR, PT, NAO, SCA, EA\_JET, NP, EP, WP

The R-squared value can be interpreted as the proportion of the variation of the predictand that is “described” or “accounted for” by the regression (Wilks, 1995). The R-squared is adjusted when there are multiple predictands, creating the adjusted R-squared coefficient. An adjusted R-squared of 0.60 (describing 60% of the predictand variance) or greater is the goal if any predictive model were to be discovered. As one can see from

Table 5, the greatest adjusted R-squared, for just one year, was 43%. The rest of the results using May teleconnections versus June-August CDDs are listed in Table 5. The results are not conducive to a predictive model, so data mining techniques were used for further exploration.

Table 5. P-value and adjusted R-squared from the ANOVA table for the 14 locations. Linear regression used May teleconnections and June-August CDD's.

City	<b>ATL</b>	<b>CHI</b>	<b>CIN</b>	<b>DFW</b>	<b>DM</b>	<b>LV</b>	<b>MEM</b>
<b>p-val</b>	<0.0001	0.085	0.005	0.001	0.110	0.002	0.002
<b>Adj R<sup>2</sup></b>	0.426	0.119	0.131	0.240	0.329	0.097	0.310
City	<b>MIN</b>	<b>NYL</b>	<b>PHI</b>	<b>POR</b>	<b>SAC</b>	<b>TUC</b>	<b>WPAFB</b>
<b>p-val</b>	0.005	0.139	0.001	0.002	0.002	0.016	0.062
<b>Adj R<sup>2</sup></b>	0.289	0.285	0.053	0.314	0.290	0.205	0.114

## **V. Tree-Based Statistical Models**

### **Overview**

Tree-based statistical models are a recent development in statistics that have been applied to prediction problems in widely diverse fields of endeavor, but are, as of yet, not well known in the atmospheric sciences (Burrows and Assel, 1992). This study uses classification and regression trees (CART) analysis to explore the data. CART is a tree-based statistical procedure for application to classification and regression problems. Breiman et al. (1984) found that error rates of CART solutions are nearly always as low or lower than solutions by linear regression. Error rates are also significantly lower for problems involving complex predictands and many predictors (Burrows and Assel, 1992).

From a database of predictand cases and accompanying predictors, CART establishes decision trees that are a classification of categorical predictands or a regression of continuous predictands. A decision tree consists of a tree-like structure of binary decisions rules. At each decision point (node) a case will branch either to the left or right based on a test against a specific predictor value, and will continue branching until a final point (terminal node) is reached. CART uses input parameters of tree length, parent node size, and child node size to determine the number of nodes. It uses the inputs to search for the tree that provides the least error when used with independent data. In this study, the independent data are represented by the ten years of data that was withheld from the original dataset. After a tree is calculated, a process of eliminating terminal nodes (pruning) is accomplished to make the tree a more effective model. Categorical

predictors are used in classification tree analysis and continuous predictors are used in regression tree analysis (Burrows and Assel, 1992).

The goal of this study at this point was to produce a predictive tool, using CART analysis, for seasonal HDDs or CDDs that would be more accurate than using the climatological normals or simple frequency distributions of occurrences.

### **Classification Tree Analysis**

Classification trees were the first tree-based models attempted. To use this model the data had to be categorized into a nominal data format. Data were categorized into thirds, using categories of above normal, normal, and below normal. Each HDD and CDD vector was sorted into ascending order, then the separation values between the upper third, the middle, third and the lower third were calculated. All data between the calculated figures in each vector were considered in the specific group of above normal, normal, or below normal categories.

An example of such a tree is shown in Figure 16. A brief explanation of this classification tree provides the reader a general idea of the tree's structure. This tree was computed using data from Minneapolis, Minnesota, using May teleconnections and categorized June-August CDDs. Specific "parent" and "child" node inputs are user provided. In this tree the parent node of any split must have at least  $n=6$ ,  $n$  being the number of data points (years) in the node, and the child node must be at least  $n=2$ . If these conditions are not met, the node will stop splitting. For example, node 5 has  $n=6$ , but the program calculated that if this node was split, one of the resulting splits would not be at least  $n=2$ . The split was therefore stopped. However, in node 4, with  $n>6$ , a split was accomplished because the child nodes were both at least  $n=2$ .



To reach a specific node, a series of conditions must be true. For example: to get to node 10 the EP must be less than or equal to 0.45, the EAWR must be greater than or equal to -0.2, and EA\_JET must be greater than -0.35.

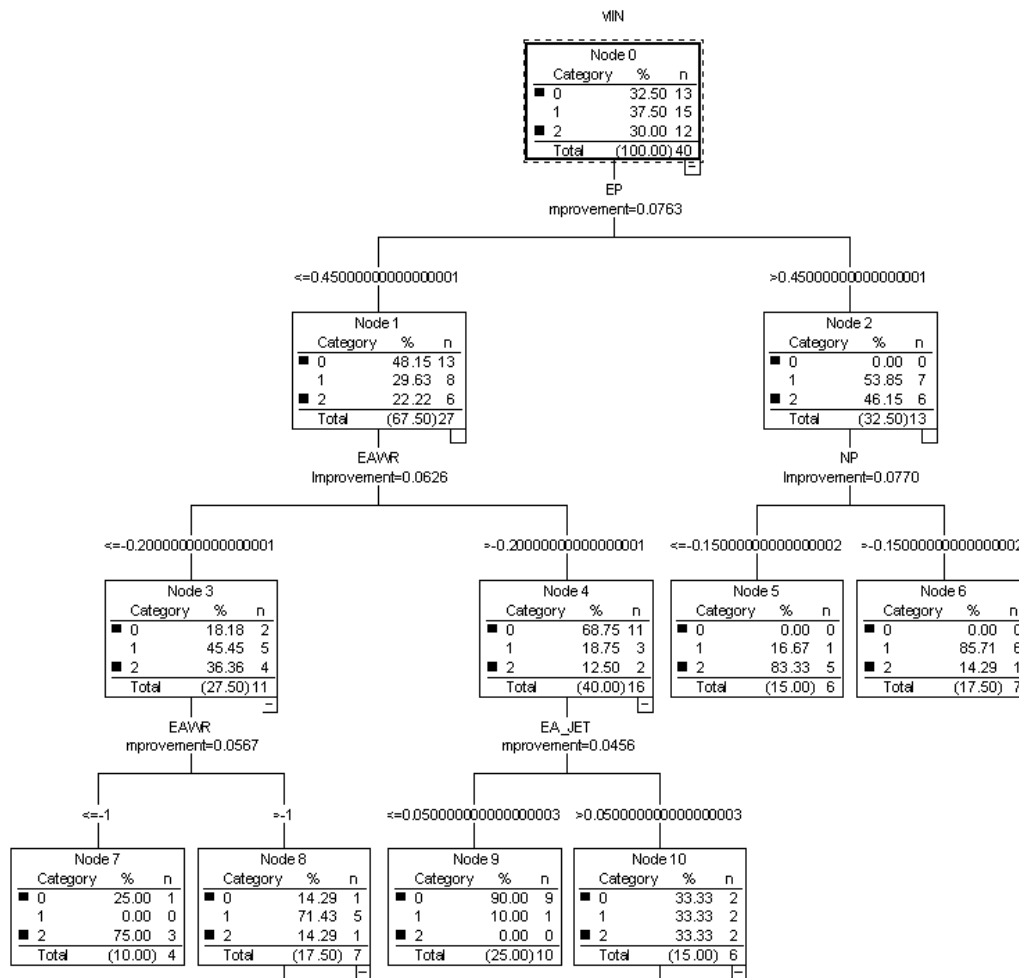


Figure 15. Example of a classification tree. This example is a tree run with data from Minneapolis using May teleconnections and June-August CDDs. Three categories are present; 2 is above normal, 1 is normal, and 0 is below normal.

Extracting any predictive information proved difficult in the classification trees. For example: node 9 shows nine remaining data points from the original 13 in the below normal category (69%), with nine of the 10 data points in that node (90%). This gives a total probability of ending in node 9 of 62% below normal, which is not a bad result. However, the best probabilities calculated from the trees were in the mid 60% range and only in a few nodes. In addition, specific conditions needed to exist to arrive in the nodes. In this example, node 9 only incorporates 25% of the total data. This result created difficulty in creating a tool that would efficiently incorporate the whole dataset. It didn't appear there was any likelihood of creating any useful predictive tools from classification trees, so a different form of CART analysis was accomplished.

### **Regression Tree Analysis**

The regression tree differs from the classification tree in that it uses continuous data instead of classified nominal data. Figure 17 is an example of such a regression tree. A brief explanation of this example tree will give the reader a general idea of the tree's structure. This tree was computed with the same data from Minneapolis, Minnesota, using May teleconnections and June-August CDDs. The user inputs three initial constraints before a tree can be grown. The inputs are maximum number of levels, minimum number of data points necessary in the parent node before a split can be performed, and the minimum number of data points in the child node before a split can be performed. In the example shown in Figure 17, the input values are 10 maximum levels of the tree, 6 minimum data points in the parent node, and a minimum of 2 data points in a child node. The regression tree starts with a beginning node, node 0. In this example, node 0 represents the summed CDDs from June-August for Minneapolis. It displays the

mean standard deviation, number of data points, and the percent of data that is in that particular node. Each parent node is split into two child nodes until the splitting is stopped by user specified inputs.

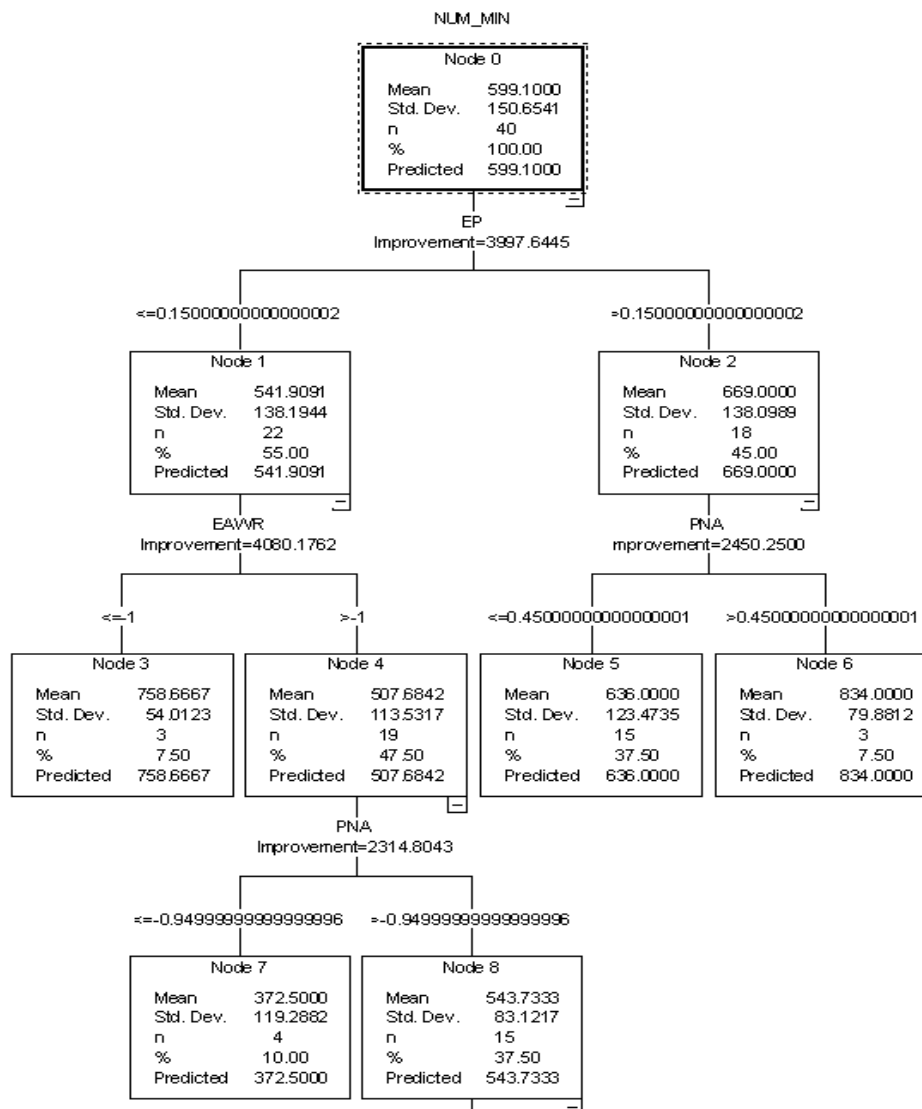


Figure 16. Example of a regression tree. Shown is a tree run with data from Minneapolis using May teleconnections and June-August CDDs.

The fundamental idea to make a split is to select each split of a node so that the data in each of the child nodes are “purer” than the data in the parent node (Breiman et al., 1984). For continuous target variables, the least-squared deviation (LSD) impurity measure is used. The LSD measure ( $R(t)$ ) is the within-node variance for node  $t$ , and is equal to the resubstitution estimate of risk for the node. It is defined as:

$$R(t) = \frac{1}{N(t)} \sum_{i \in t} (y_i - \bar{y}(t))^2 \quad (5)$$

where  $N(t)$  is the number of cases in the node  $t$ ,  $y_i$  is the value of the target variable (location HDDs or CDDs), and  $\bar{y}_i$  is the mean for node  $t$ . The LSD criterion function for split  $s$  at node  $t$  is defined as:

$$\phi(s, t) = R(t) - p_L R(t_L) - p_R R(t_R) \quad (6)$$

where  $p_L$  is the proportion of cases in  $t$  sent to the left child node,  $p_R$  is the proportion sent to the right child node, and  $t_L$  and  $t_R$  are the nodes created by the split  $s$ .

The software runs all possible splits on the node and splits the node at the location of the largest decrease in impurity. This value is shown on the tree as the “improvement”. The process is then repeated at each node (SPSS, 2001).

### **Application of Regression Tree Analysis**

The goal of this study was to come up with a predictive tool for HDDs/CDDs using teleconnection indices. To test the process at 14 locations, May teleconnections and June-August CDDs were used. First, a goodness of fit test for normality was accomplished on the CDDs in each city using the Shapiro-Wilk test. The results are shown in Table 6.

Table 6. Shapiro-Wilk goodness of fit test for normality. A p-value > 0.05 shows a normal distribution. DFW and TUC did not pass the test.

City	ATL	CHI	CIN	DFW	DM	LV	MEM
<b>W-S test</b>	0.99	0.29	0.26	<b>0.01</b>	0.42	0.55	0.31
City	MIN	NYL	PHI	POR	SAC	TUC	WPAFB
<b>W-S test</b>	0.84	0.22	0.69	0.25	0.93	<b>0.004</b>	0.12

A p-value > 0.05 in the Shapiro-Wilk test indicates a normal distribution. Dallas, and Tucson did not pass the normality test, however, only one data point for Tucson and two for Dallas created a non-normal distribution, so the exploration for a predictive outcome continued with normality assumed for all locations. With the goal of coming up with a predictive tool that is better than the climatological normals or simple frequency distributions in mind, it was decided to create a 95% prediction interval to create a range of CDDs. The mean and standard deviation from each tree node was used to create a 95% prediction interval which is defined as:

$$\bar{x} \pm t_{0.025, n-1} \cdot s \sqrt{1 + \frac{1}{n}} \quad (7)$$

where  $\bar{x}$  is the mean from the calculated node,  $t$  is the critical value for a t-distribution,  $n$  is the number of data points in the node, and  $s$  is the sample standard deviation.

The next step in tree-structured statistics is to prune the tree. Pruning consists of eliminating the terminal nodes necessary to create the best effective tree. How to prune a tree depends on the data being analyzed. The pruning criteria for the trees in this study

were calculated during the verification of the process created. The independent data (ten years) withheld from the original data were run through the trees. The teleconnections for each year were run through the tree to calculate which node was the terminating node, then the specific year's CDDs were checked to see if they fell within the created prediction interval from the same node. During the verification process the data were run through multiple models with different criteria for pruning the terminal nodes. It was found that a node with  $n < 6$  needed to be pruned. As shown in Figure 18, those nodes of the tree with  $n < 6$  are terminated.

Results of model verification are shown in Table 7. Verification results for the individual locations were between 80% and 100% with an overall 88% verification rate.

Table 7. Percentage of CDDs that were in the predicted range after verification data were run through the trees. An overall verification rate of 88% was achieved.

City	ATL	CHI	CIN	DFW	DM	LV	MEM
<b>% CDDs in final node</b>	80	87.5	80	100	90	80	80
City	MIN	NYL	PHI	POR	SAC	TUC	WPAFB
<b>% CDDs in final node</b>	90	80	90	100	100	90	90

### **Results vs. Frequency Distribution**

The goal of this study was to create a predictive tool that was better than the climatological standard normals or simple frequency distributions of CDDs/HDDs occurrences. Table 8 shows comparisons of the simple frequency distributions of occurrences of Minneapolis June-August CDDs and the created 95% prediction interval for valid nodes of the tree shown in Figure 18. The new calculated forecast ranges are

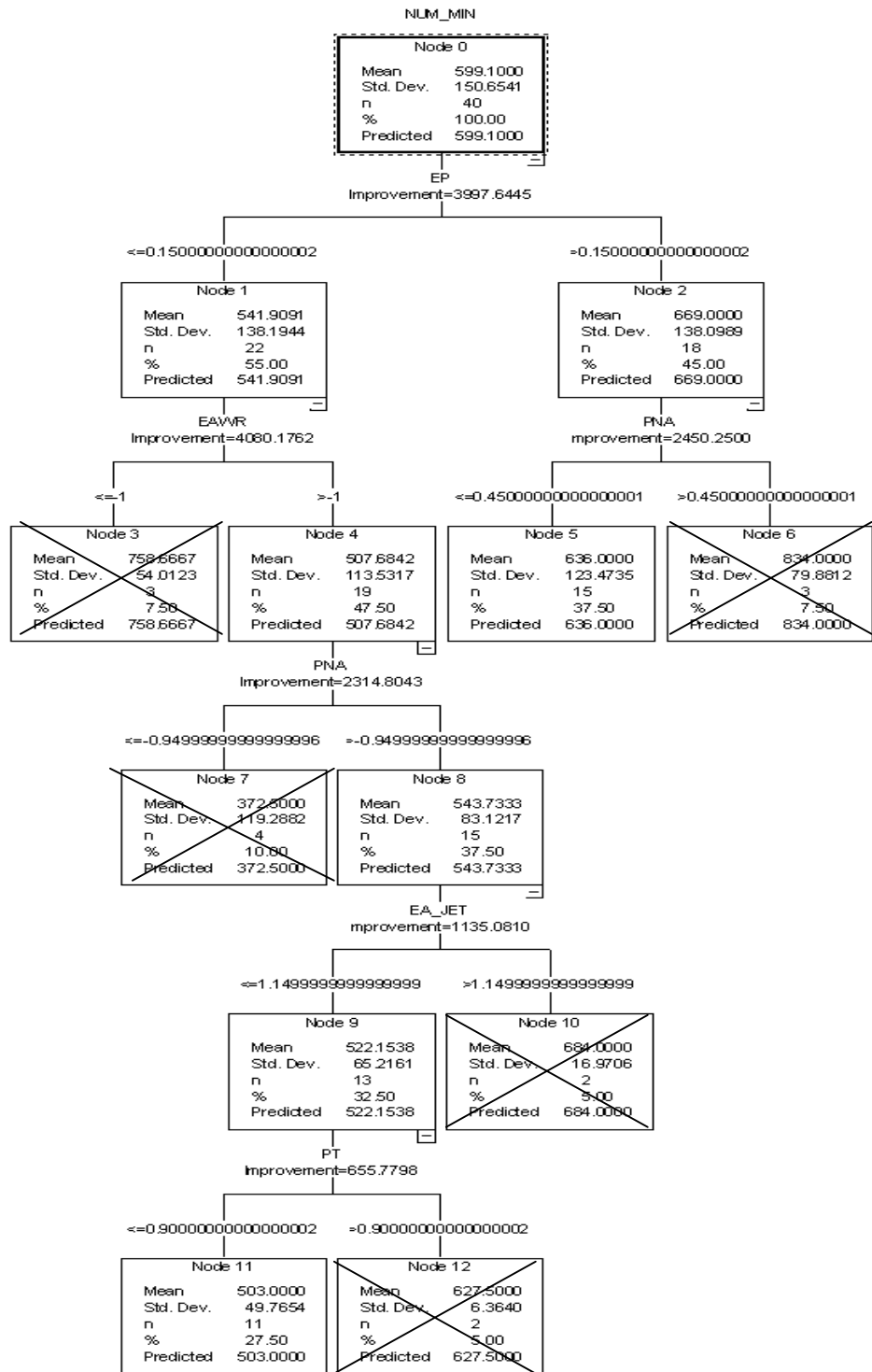


Figure 17. Example of a pruned regression tree. This example tree was run with data from Minneapolis using May teleconnections and June-August CDDs. The pruned branches are crossed out if  $n < 6$  in any node.

shaded in gray. The reduction of the original CDD range is quantified by calculating a ratio of the new forecast range, for the individual nodes, to the original CDD range and subtracting this value from one. This reduction percentage is multiplied by the climatological frequency distribution for each individual node to obtain an expected forecast range reduction percentage. The individual node expected forecast range reduction percentages are summed to obtain a total expected range reductions. This reduction percentage can be viewed as the total expected forecast range reduction over climatology.

As an example from Table 8: the total range of summed CDDs are broken into 14 ranges between 248 and 909, with the frequency distributions of occurrences for the CDD ranges in the next column. The calculated percentage the range is reduced in node 11 is the ratio of the new forecast range ( $626-390 = 236$  CDDs) with the total range (661 CDDs). Therefore, in node 11 the range is reduced ( $1-(236/661)$ ) or 64%. Climatologically, over the 40-year period of record, the calculated CDDs are in node 11 12.5% of the time. The product of the reduced range (64%) and the observed climatological frequency of occurrences per node (12.5%) shows an expected forecast range reduction for node 11 of 8% over climatology. Summing all of the individual node expected forecast range reduction percentages shows a total expected forecast range reduction of 36.45% over climatology for Minneapolis. The expected forecast range reduction for all 14 locations are shown in Table 9, with an overall expected forecast range reduction of 35.7% over climatology. This value varies from 16.8% for Cincinnati to 58.9% for Tucson.



Table 8. Expected forecast range reduction. Prediction intervals computed in each node are shown in gray. The percentage the range is reduced is multiplied by the climatological frequency distribution to obtain an expected forecast range reduction. The individual node expected forecast range reduction percentages are summed to obtain a total expected forecast range reduction over climatology, 36.45% in this case.

Node										
MIN CDD	Frequency Distribution	0	1	2	4	5	8	9	11	14
>248<=295	0.0250	0	0	0	0	0	0	0	0	0
>295<=342	0.0000	1	1	0	1	0	0	0	0	0
>342<=390	0.0500	1	1	0	1	0	0	0	0	0
>390<=437	0.0250	1	1	1	1	1	1	1	1	0
>437<=484	0.1500	1	1	1	1	1	1	1	1	1
>484<=531	0.0750	1	1	1	1	1	1	1	1	1
>531<=579	0.2000	1	1	1	1	1	1	1	1	1
>579<=626	0.0750	1	1	1	1	1	1	1	1	1
>626<=673	0.1000	1	1	1	1	1	1	1	0	1
>673<=720	0.1000	1	1	1	1	1	1	0	0	0
>720<=767	0.0750	1	1	1	1	1	1	0	0	0
>767<=815	0.0250	1	1	1	0	1	0	0	0	0
>815<=862	0.0500	1	1	1	0	1	0	0	0	0
>862<=909	0.0250	1	0	1	0	1	0	0	0	0
>909	0.0250	0	0	1	0	1	0	0	0	0
Percent reduction in forecast range per individual node (%)		7	14	22	29	22	43	57	64	64
Climatological frequency distribution of occurrences per node (40 years)		0.000	0.075	0.075	0.100	0.375	0.050	0.050	0.125	0.150
Expected forecast range reduction. (%)		0.00	1.05	1.65	2.90	8.25	2.15	2.85	8.00	9.60
Total										36.45

Table 9. The total expected forecast range reduction over climatology for the 14 forecast locations. The *overall average* expected forecast range reduction for this example is 35.7% over climatology.

City	ATL	CHI	CIN	DFW	DM	LV	MEM
<b>Expected forecast range reduction (%)</b>	39.7	33.3	16.8	32.2	48.2	46.4	34.5
<b>City</b>	MIN	NYL	PHI	POR	SAC	TUC	WPAFB
<b>Expected forecast range reduction (%)</b>	36.5	39.4	36.1	24.9	30.2	58.9	22.7

## **VI. Conclusions and Recommendations**

### **Conclusions**

This study has introduced a new technique to significantly increase the accuracy of seasonal long-range temperature forecasts. It statistically explored teleconnection indices and, using a tree-based statistical regression, created a predictive tool for future CDDs and HDDs summed over three months.

Temperature data were gathered from 14 U.S. locations in order to represent most of the climate regimes across the country (Objective 1). HDDs and CDDs were calculated using the temperature data gathered to use as predictor variables (Objective 2). Teleconnection indices from the 13 most significant Northern Hemispheric teleconnections and the Southern Oscillation Index in the Southern Hemisphere were gathered to use as predictand variables (Objective 3). Ten years of data were then removed for independent verification of the technique created (Objective 4).

Linear regression analysis was accomplished on the data using teleconnections from May and summed CDDs from June-August. Valid models were found during the analysis, but the amount of variance of the predictand explained by the linear regression was rarely greater than 35%, in which case, creating a predictive tool would be difficult (Objective 5).

Tree-based analysis was accomplished on the data (Objective 6), first using classification tree analysis; however, extracting any predictive information also proved difficult with this type of approach. Regression tree analysis was then accomplished on

the data. Trees were created and the predicted mean and standard deviations were used to create a method for predicting seasonal CDDs and HDDs.

This new technique creates a range of HDDs/CDDs that is significantly more accurate than simple frequency distributions of occurrences. The predicted mean and standard deviations from the regression tree output were used to calculate 95% prediction intervals for each of the nodes. Teleconnections were run through the trees to compute a predicted node, and then the new interval for the predicted node was used as the predictive range for the HDDs/CDDs for the particular forecast months (Objective 7).

This new model verified, using 10 years of independent data withheld from the original data set (Objective 8), at an excellent 88% overall verification rate with 3 of the 12 cities verifying at 100%. Two other cities, which verified at the 90% significance level, failed in the randomly selected year of 1988. This year is a well-known El Nino year and record temperatures were experienced in some parts of the U.S. The summed CDDs for Minneapolis and WPAFB in 1988 fell outside the range of the original data set. Extrapolation of the model to fit the data outside the range of the original data set was not accomplished because the new fitted relationship may not have been valid for such outlier values. Had the numbers for 1988 been in the original data set, the results may have been even better than they were with possibly two more cities verifying at 100%.

An expected range reduction percentage over climatology was created from the calculated ranges. An average expected forecast range reduction percentage of 35.7% was found in this study.

The question of spatial homogeneity arose during this study, but the scope of this study could not focus on the aspect of spatial homogeneity. However, because WPAFB

was included in the study, two cities, Cincinnati and WPAFB, were in close enough spatial proximity of each other to investigate the spatial homogeneity of the created prediction trees. Cincinnati verification data were run through the WPAFB trees and the results were comparable to the WPAFB results. Additionally, WPAFB verification data were run through the Cincinnati trees and the results were comparable to the Cincinnati results. These results show spatial connections between the computed trees (Objective 9).

Overall this study attempted to improve upon the methods currently used to produce long-range forecasts of temperature over the U.S. Excellent results were achieved and predictive tree tools were created which are deemed ready for users to use now for long-range temperature forecasts. It is the conclusion of this study that this innovative method works. It is also concluded that this method may be used to predict multiple atmospheric variables, well in advance, for most locations within the Northern Hemisphere.

## **Recommendations**

This study created a new technique in the way we can analyze atmospheric parameters. The hope is that this study will be a stepping-stone to future research to fully understand the magnitude of this type of analysis. Continuation of research on this study should be according to the following:

1. Try to understand how the regression tree analysis relates to physical atmospheric synoptic circulation patterns. Understanding further why the regression tree analysis splits where it does and why it uses the teleconnections in the order that it does.

2. Try to extract the model from the software in order to fully automate the technique. The teleconnections are currently run through the trees manually to calculate the terminating node. The new prediction intervals are created after the data calculated in the trees nodes are manually entered into a statistical spreadsheet. The complete process needs to be automated.

The research opportunities using this process are limitless. Currently DoD is looking for long-range seasonal forecasts for parameters over Afghanistan. This method could be used anywhere in the Northern Hemisphere. This method could also be used to predict any parameter, to include the vital ones necessary in a wartime scenario, such as cloud cover, precipitation, and visibility. This information could revolutionize long-range prediction efforts to help with humanitarian aid operations for the timing and movement of supplies.

## Bibliography

- Anthes, R. A., 1986: The General Question of Predictability, *Mesoscale meteorology and forecasting*, P.S. Ray, Ed, American Meteorology Society, 636-656.
- Barnston, A. G., and R. E. Livezey, R. E. 1987: Classification, Seasonality and Persistence of Low-Frequency Atmospheric Circulation Patterns. *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.
- Breiman et al., 1984: *Classification and Regression Trees*, Wadsworth International Group, 358pp.
- Burroughs, W. J., 1992: *Weather Cycles Real or Imaginary?* Cambridge University Press, 201pp.
- Burrows W. R. and R.A. Assel, 1992: Use of CART for Diagnostic and Prediction Problems in the Atmospheric Sciences, *12<sup>th</sup> Conference on Probability and Statistics in the Atmospheric Sciences*, American Meteorology Society, 161-166.
- CPC, 2001: <http://www.cpc.ncep.noaa.gov/data/teledoc/telecontents.html>.
- Gleick, J. 1987: *Chaos*. Penguin books, 352pp.
- Glickman T. S., Editor, 2001: *Glossary of Meteorology*, American Meteorological Society, 855pp.
- Horel, J. D., 1981: A Rotated Principal Component Analysis of the Interannual Variability of the Northern Hemisphere 500mb Height Field, *Monthly Weather Review*, 2080-2092.
- Horan, F., 1997: How much did the Gulf War cost the U.S?, <http://comp9.psych.cornell.edu>.
- Lowther, R. P., 1998: *The Development of a Seasonal Climate Forecast Methodology for ITCZ Associated Rainfall Applied to Eastern Africa*. PhD dissertation. Texas A&M University, College Station TX, 127pp.
- Mapquest.com, 2001: <http://www.mapquest.com>.

- Robinson and Henderson-Sellers, 1999: *Contemporary Climatology*. Pearson Education Limited, 317pp.
- Rodionov and Assel, 2000: Atmospheric Teleconnection Patterns and Severity of Winters in the Laurentian Great Lakes Basin, *Atmoshpere-Ocean*, XXXVIII, 601-635.
- SPSS, 2001: *AnswerTree 3.0 User's Guide*, SPSS Inc. 226pp.
- USAF, 1991: Message Number 071659ZJAN91.
- Van Loon, H. and J. C. Rogers, 1979: The seesaw in winter temperatures between Greenland and Northern Europe, Part II: Some ocean atmospheric effects in middle and high latitudes, *Monthly Weather Review*, 107, 509-519.
- Walker G. T. and E. W. Bliss, 1932: World Weather, *Memoirs of the Royal Meteorological Society*. 4: 53-84.
- Wallace, J. M. and D. S. Gutzler, 1981: Teleconnections in the Geopotential Height Field During the Northern Hemisphere Winter, American Meteorological Society, *Monthly Weather Review*, 109, 784-812.
- Washington, R., et al., 2000: Northern Hemisphere Teleconnection Indices and the Mass Balance of Svalbard Glaciers, *International Journal Climatology*, 20, 473-487.
- Wilks D. S. 1995: *Statistical Methods in the Atmospheric Sciences*, Academic Press, 467pp.



## **Appendix: Regression Trees**

The appendix contains the regression trees used in this study, which can be used as a predictive tool. They were created using May teleconnection indices and summed CDDs for June-August. Using the predicted mean and standard deviation, prediction intervals are made for each valid node ( $n > 5$ ). Overall, this prediction interval is 90% likely to contain the predicted HDDs/CDDs for the upcoming June-August with a 35.7% overall decrease in expected range over climatology.



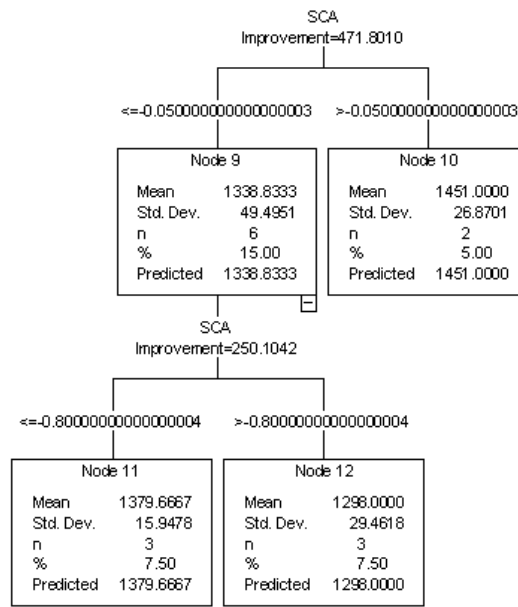


Figure 18 (continued). Atlanta regression tree.

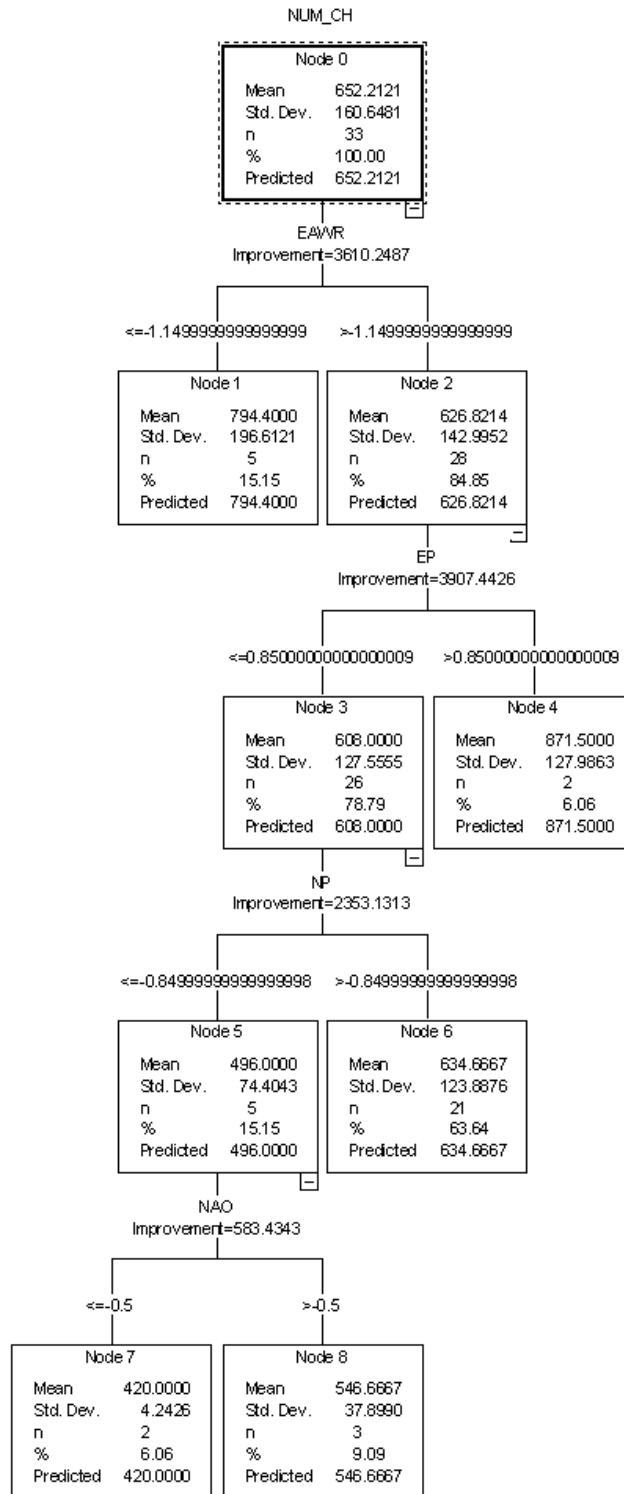


Figure 19. Chicago regression tree.

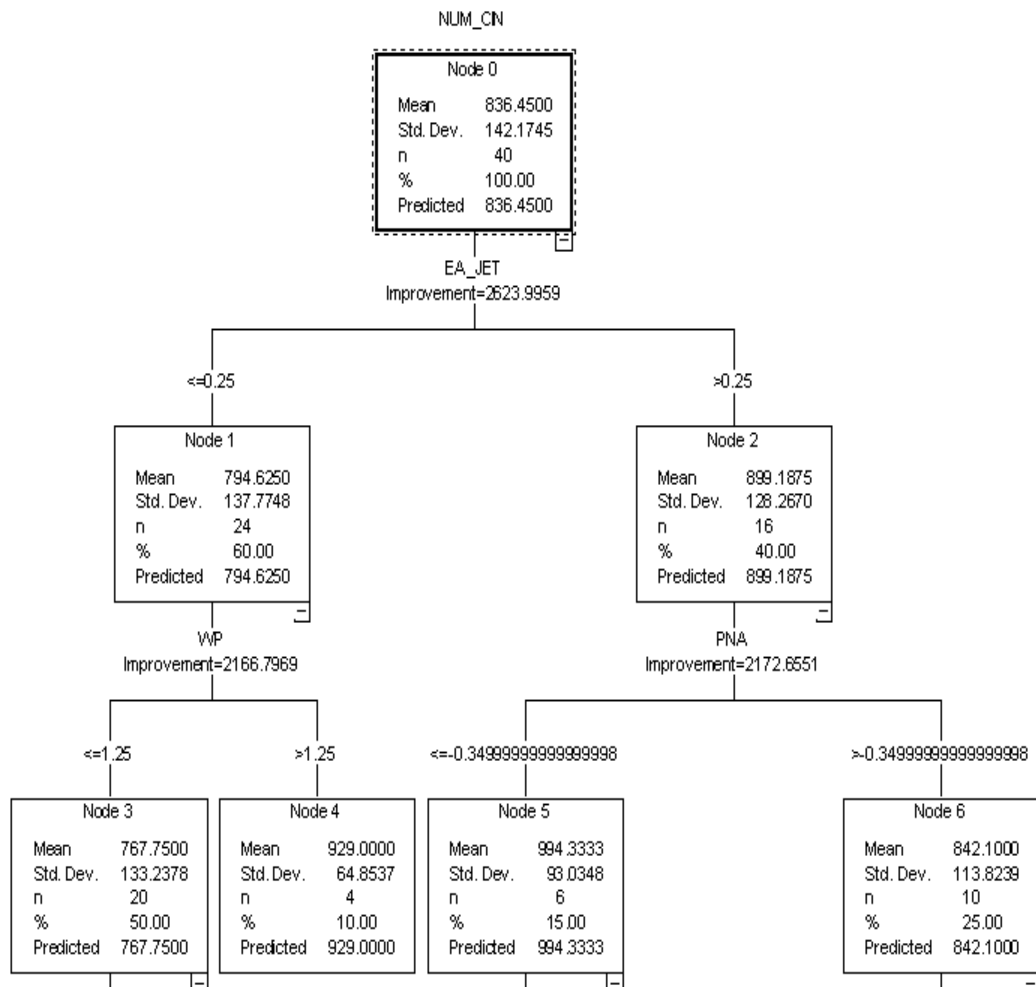


Figure 20. Cincinnati regression tree.

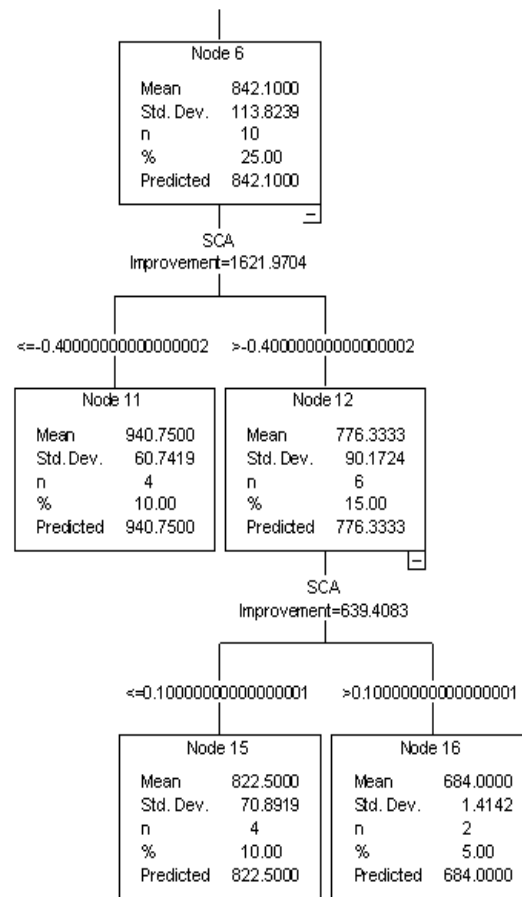


Figure 20 (continued). Cincinnati regression tree.

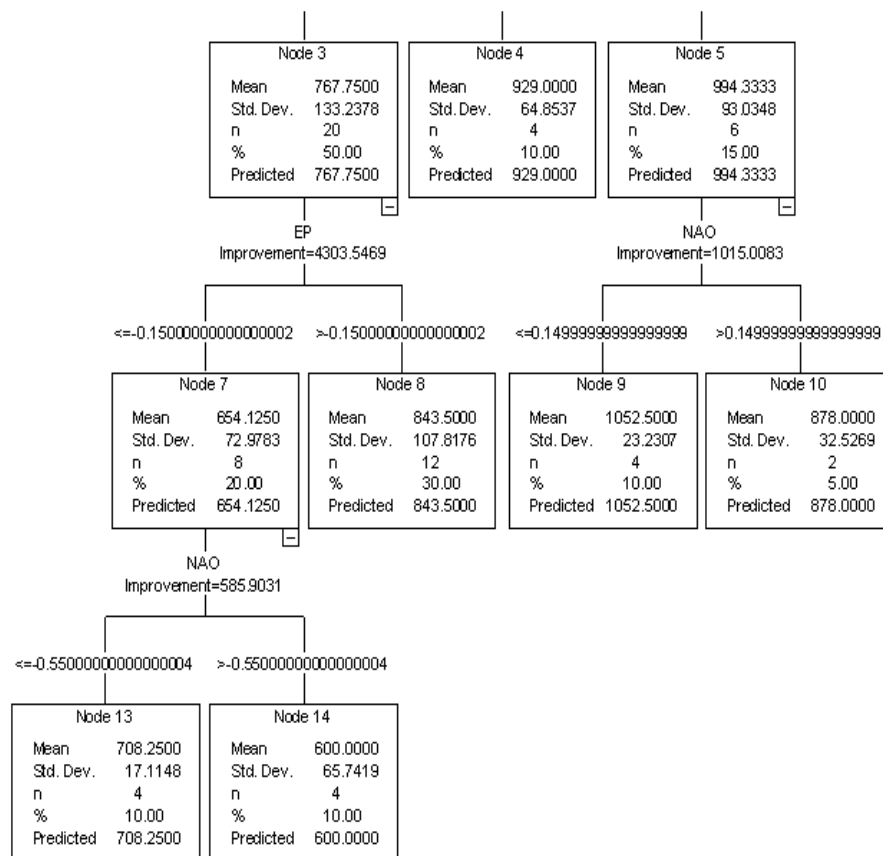


Figure 20 (continued). Cincinnati regression tree.

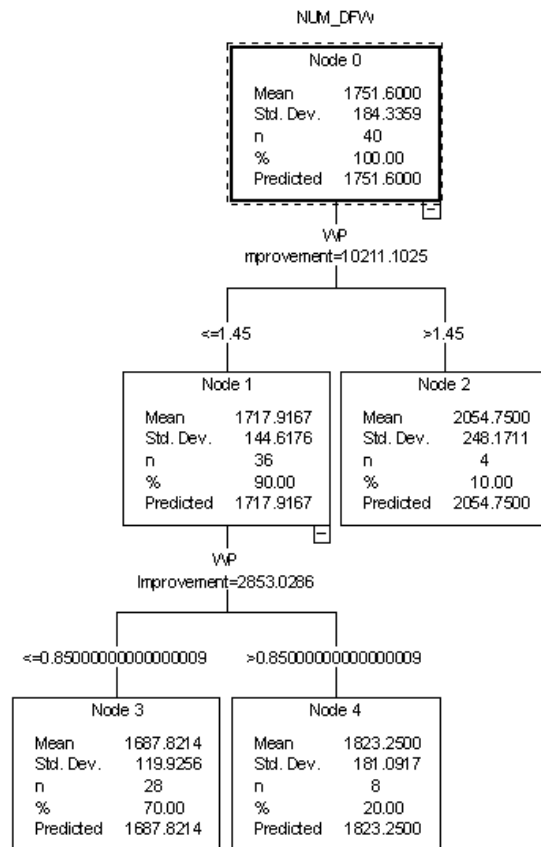


Figure 21. Dallas-Fort Worth regression tree.



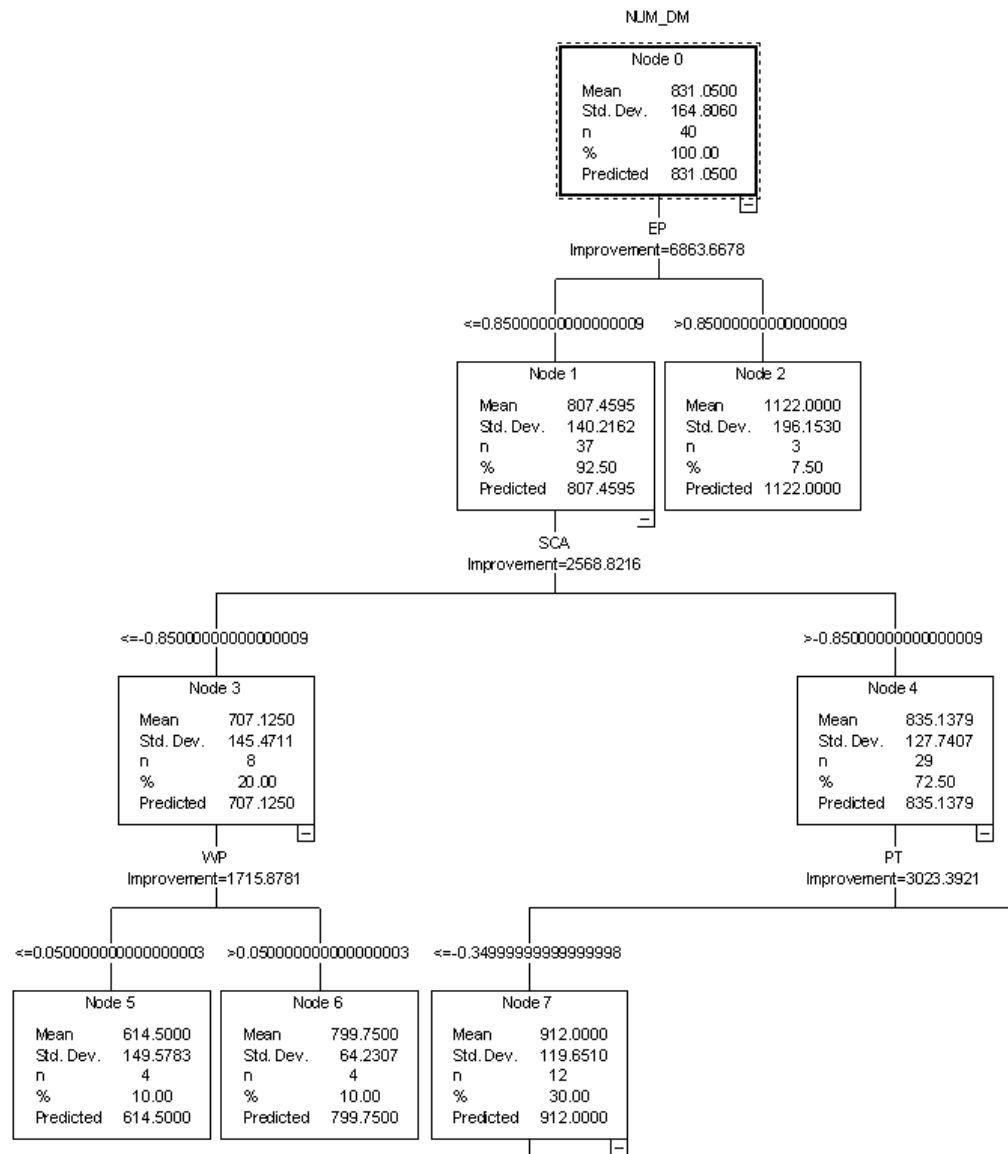


Figure 22. DeMoines regression tree.

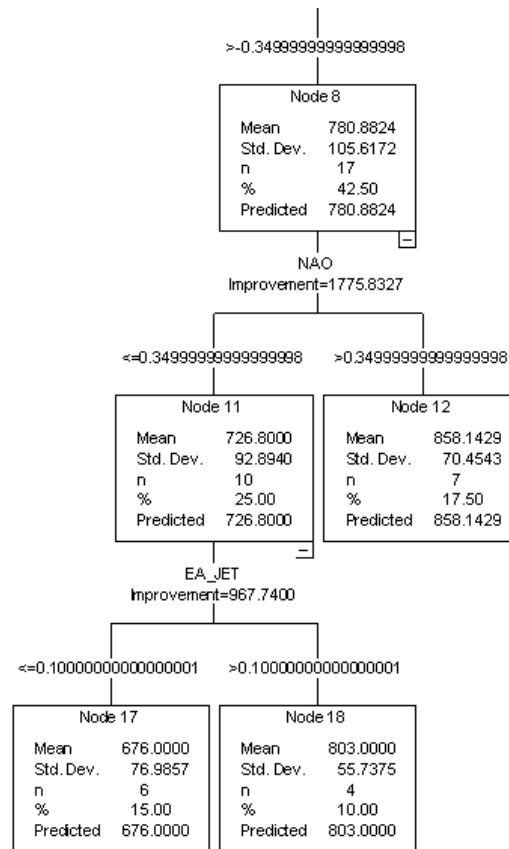


Figure 22 (continued). DeMoines regression tree.

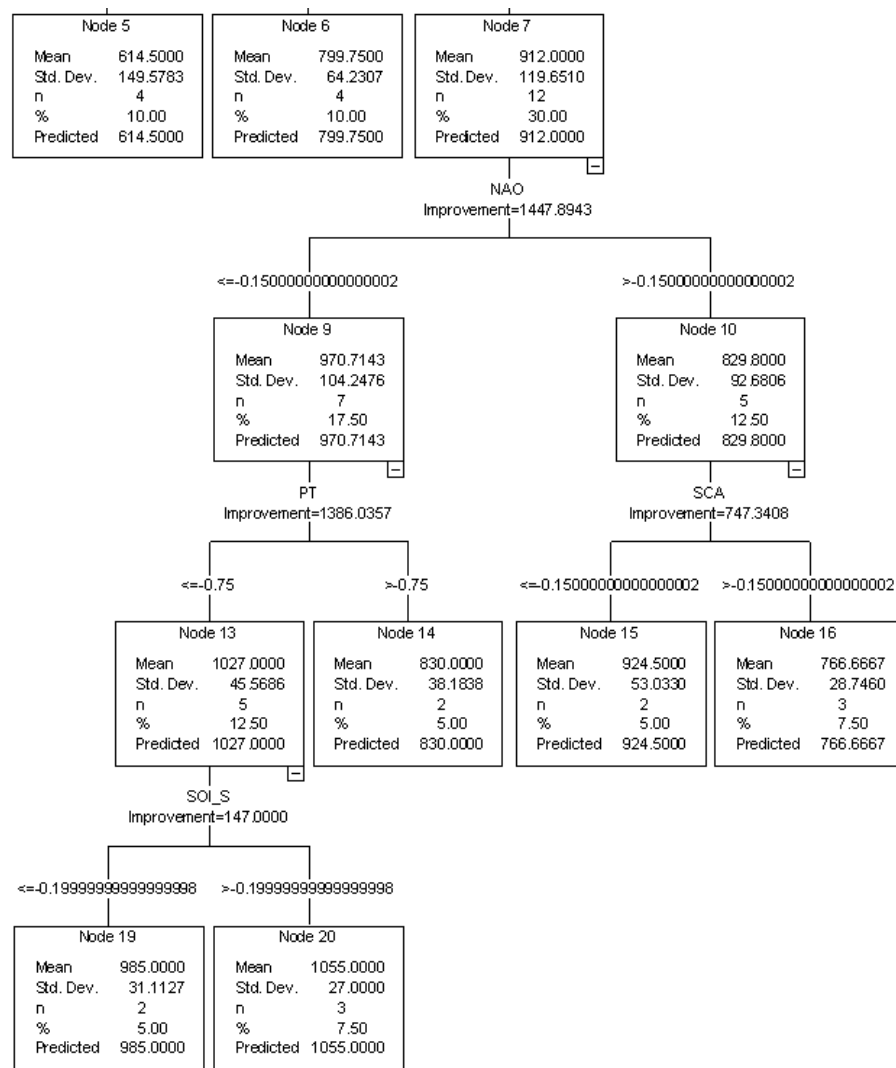


Figure 22 (continued). DeMoines regression tree continued.

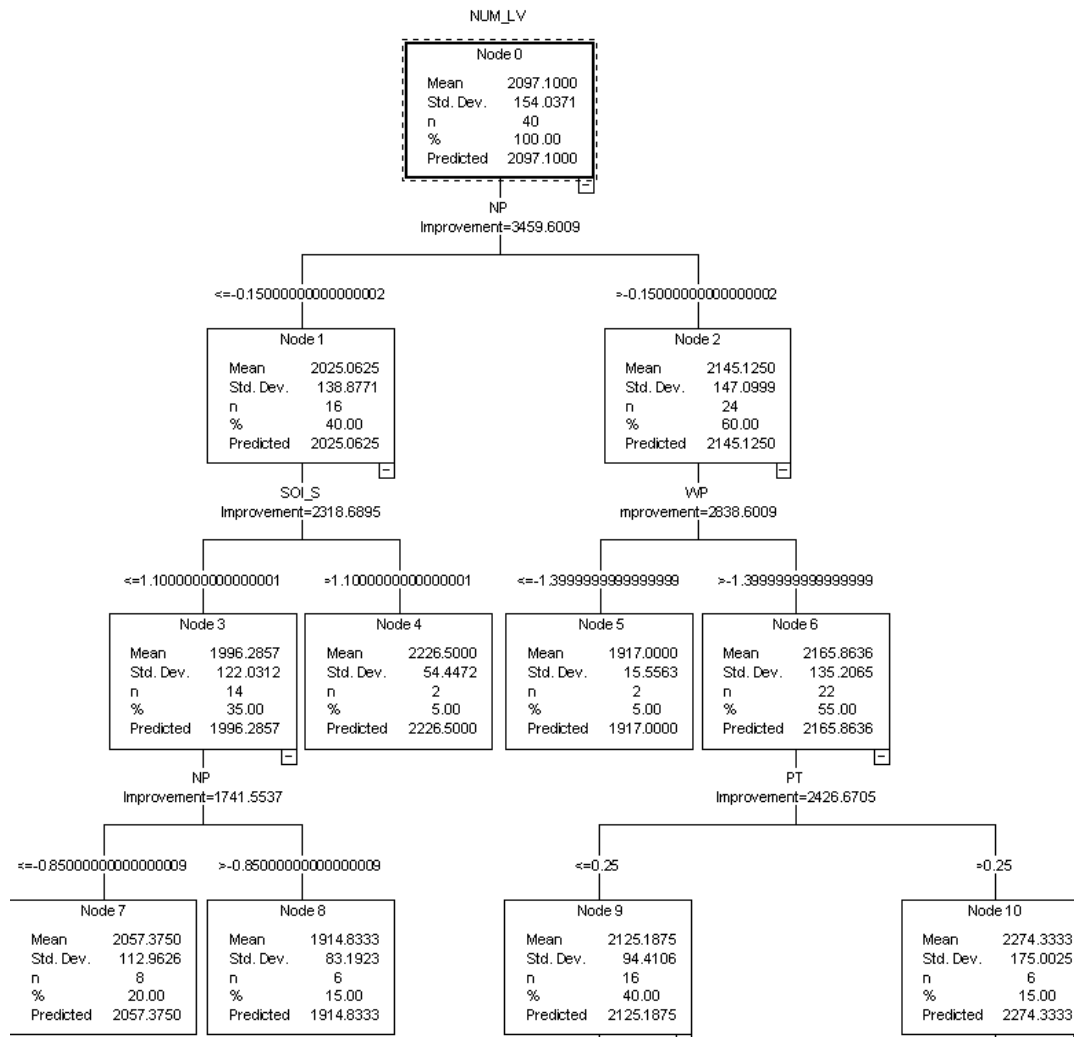


Figure 23. Las Vegas regression tree.

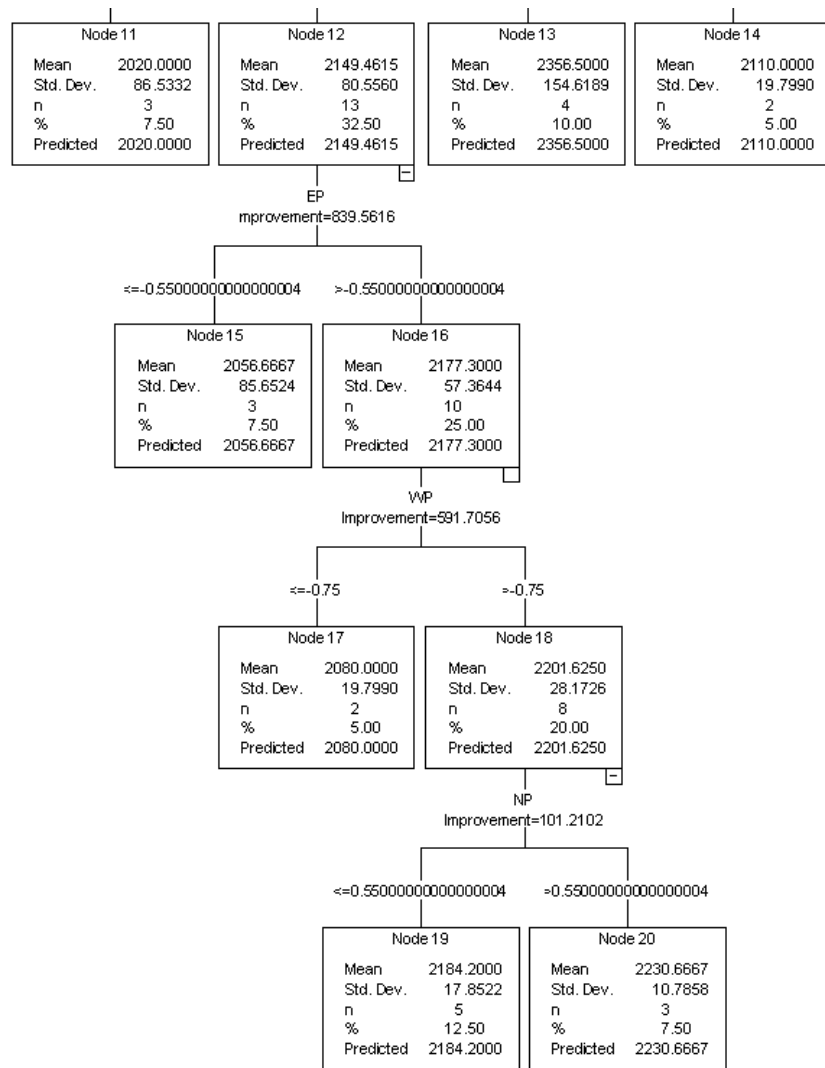


Figure 23 (continued). Las Vegas regression.

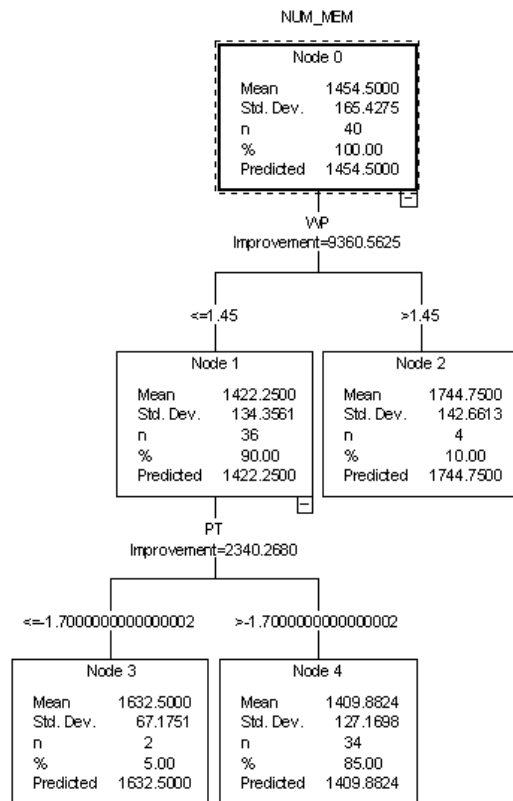


Figure 24. Memphis regression tree.

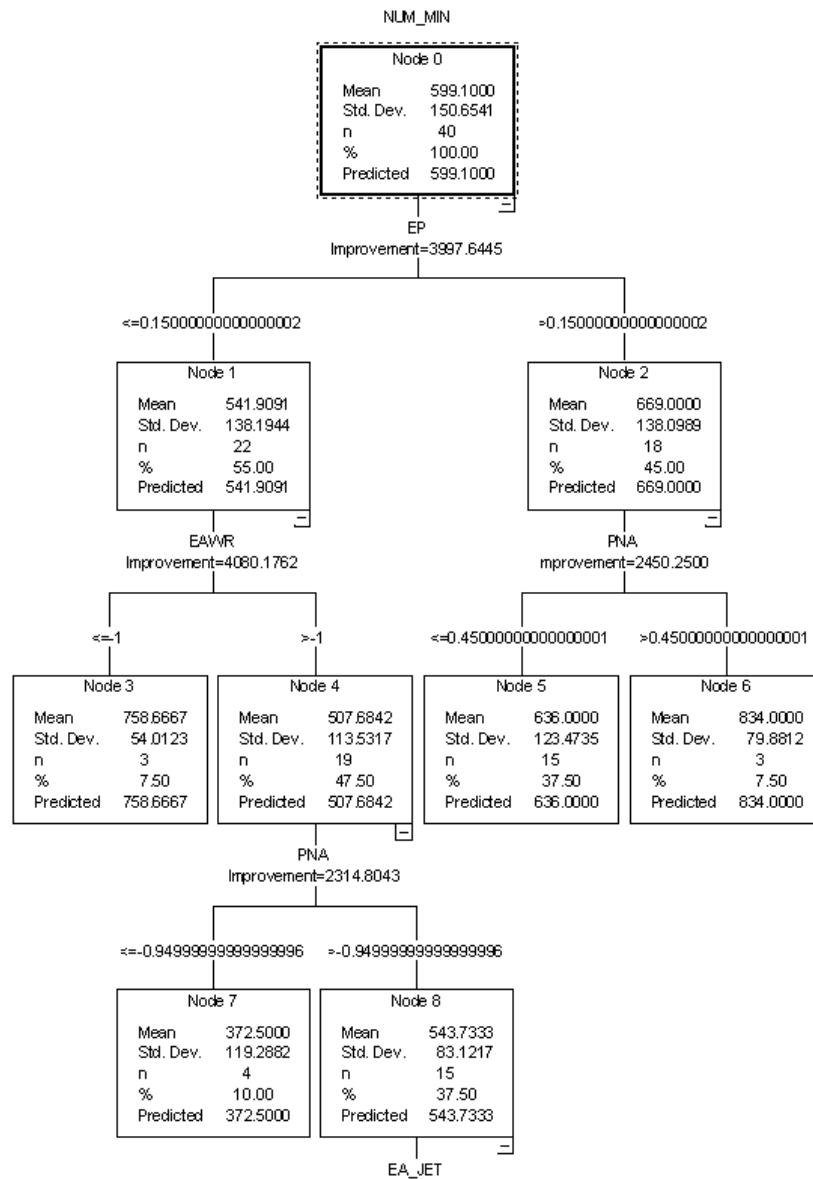


Figure 25. Minneapolis regression tree.

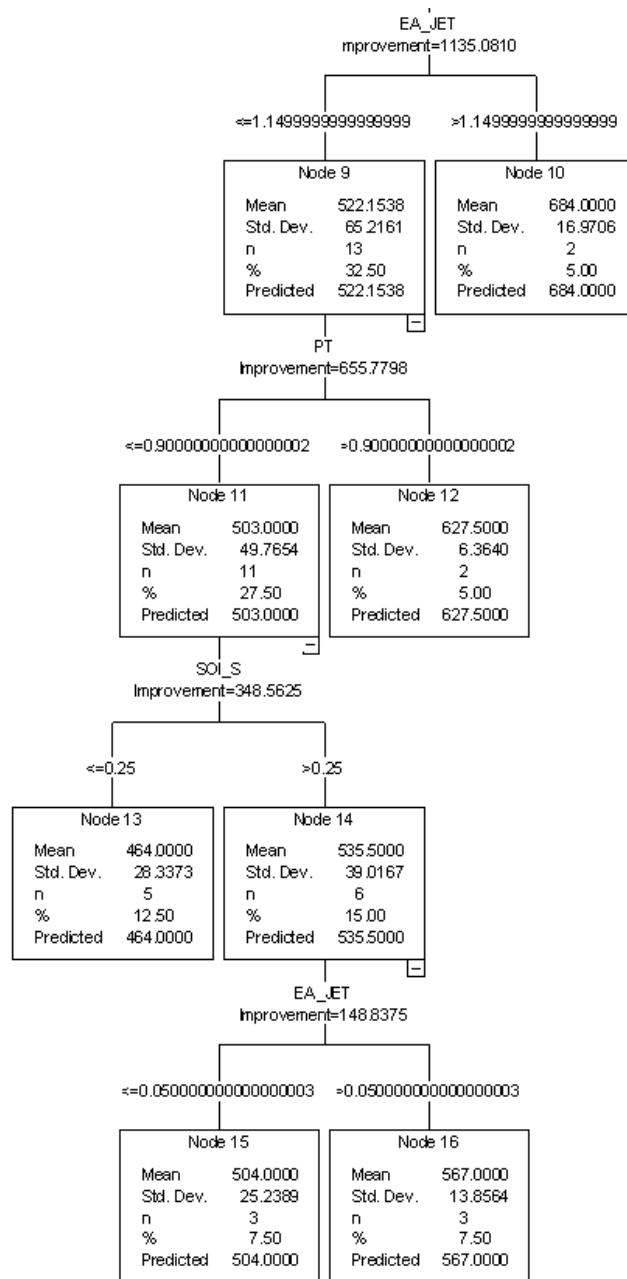


Figure 25 (continued). Minneapolis regression tree.



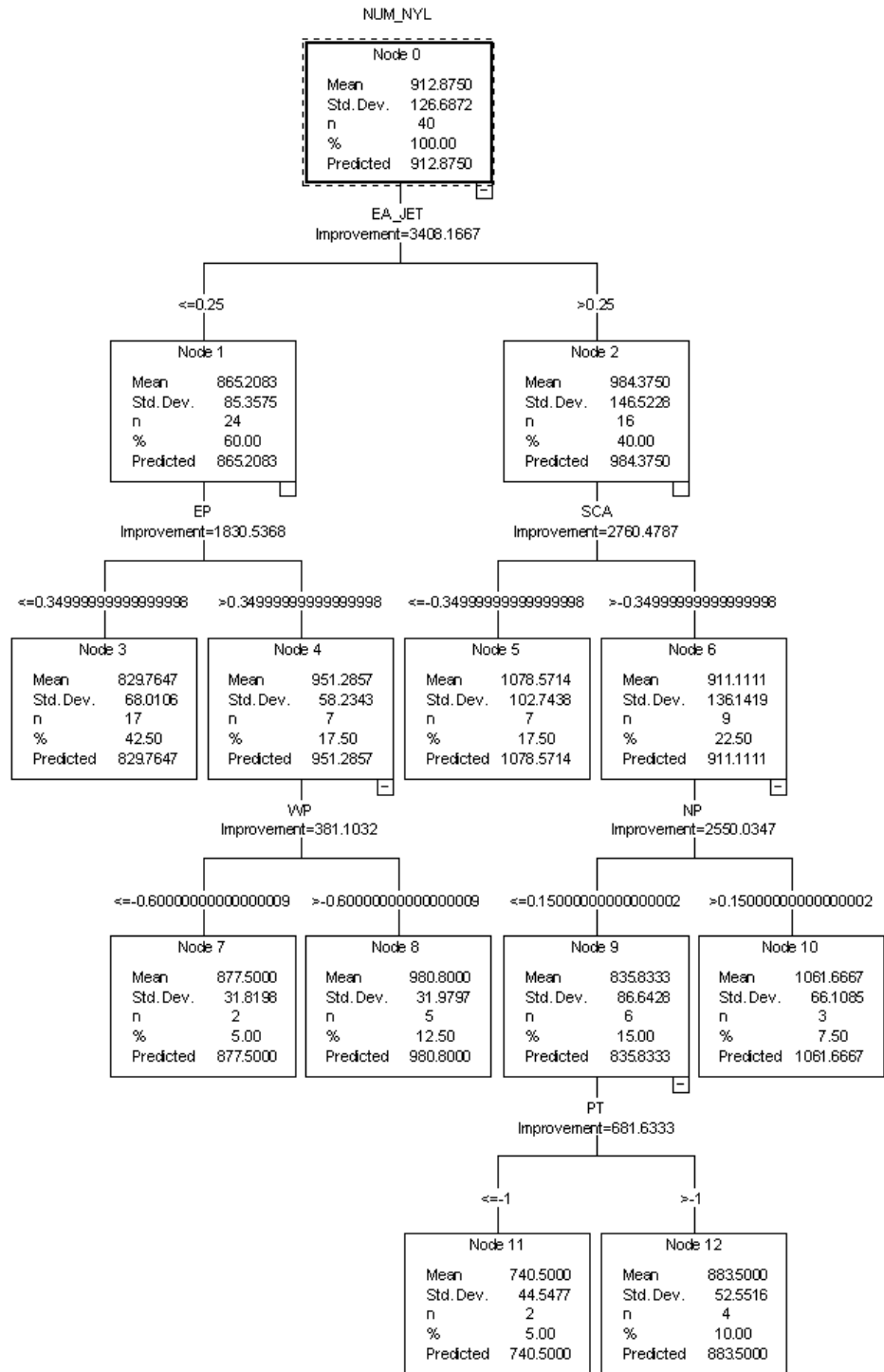


Figure 26. New York, LaGuardia regression tree.

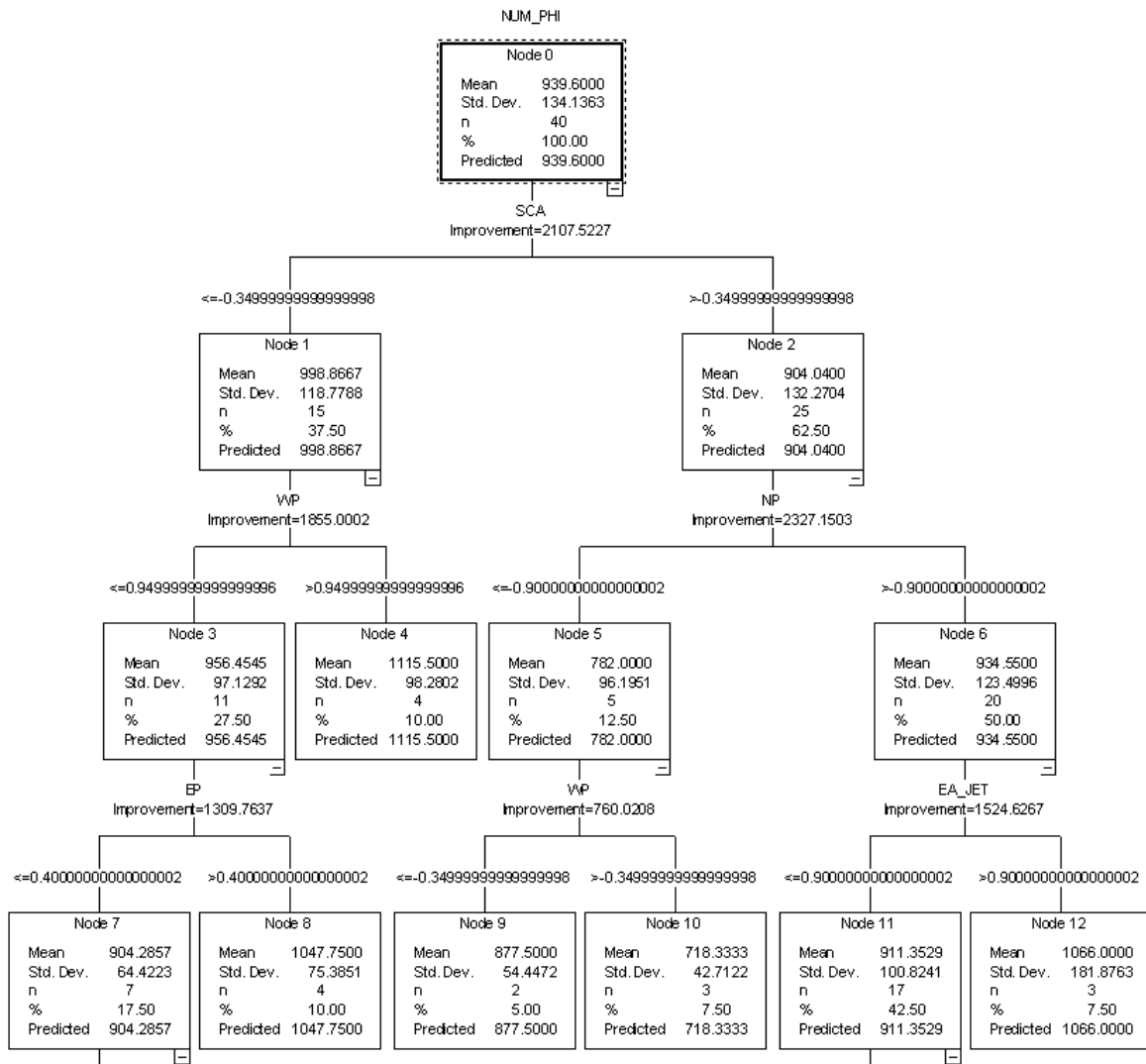


Figure 27. Philadelphia regression tree.

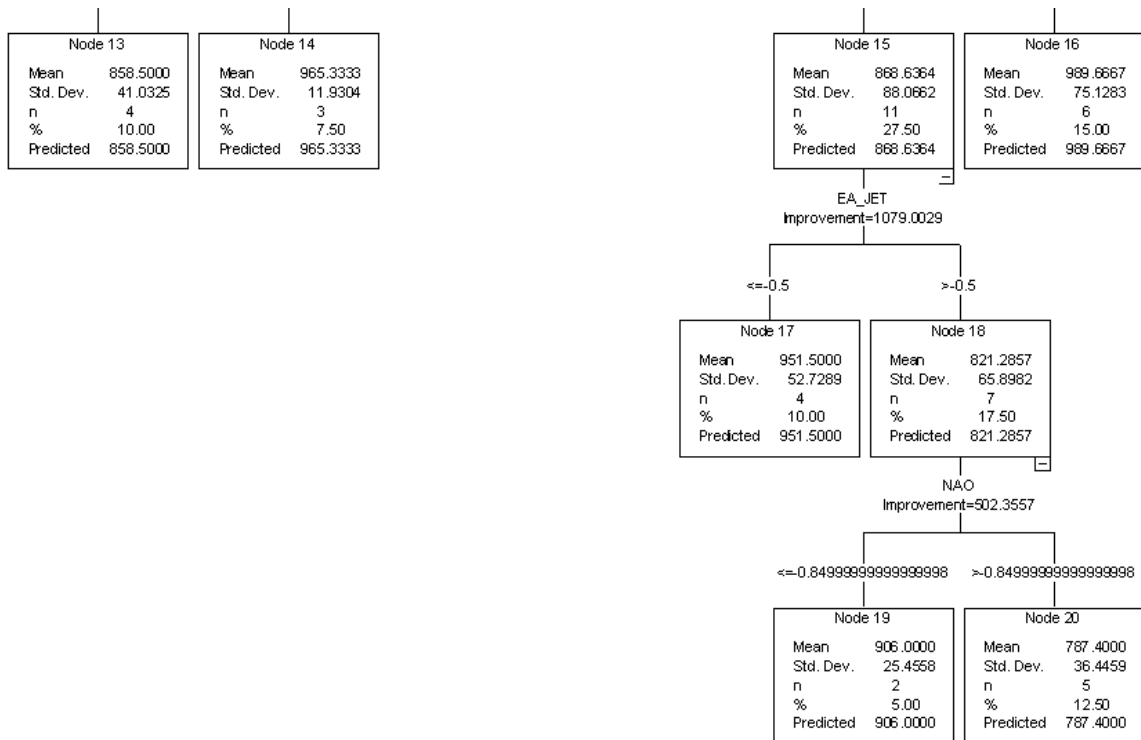


Figure 27 (continued). Philadelphia regression tree.

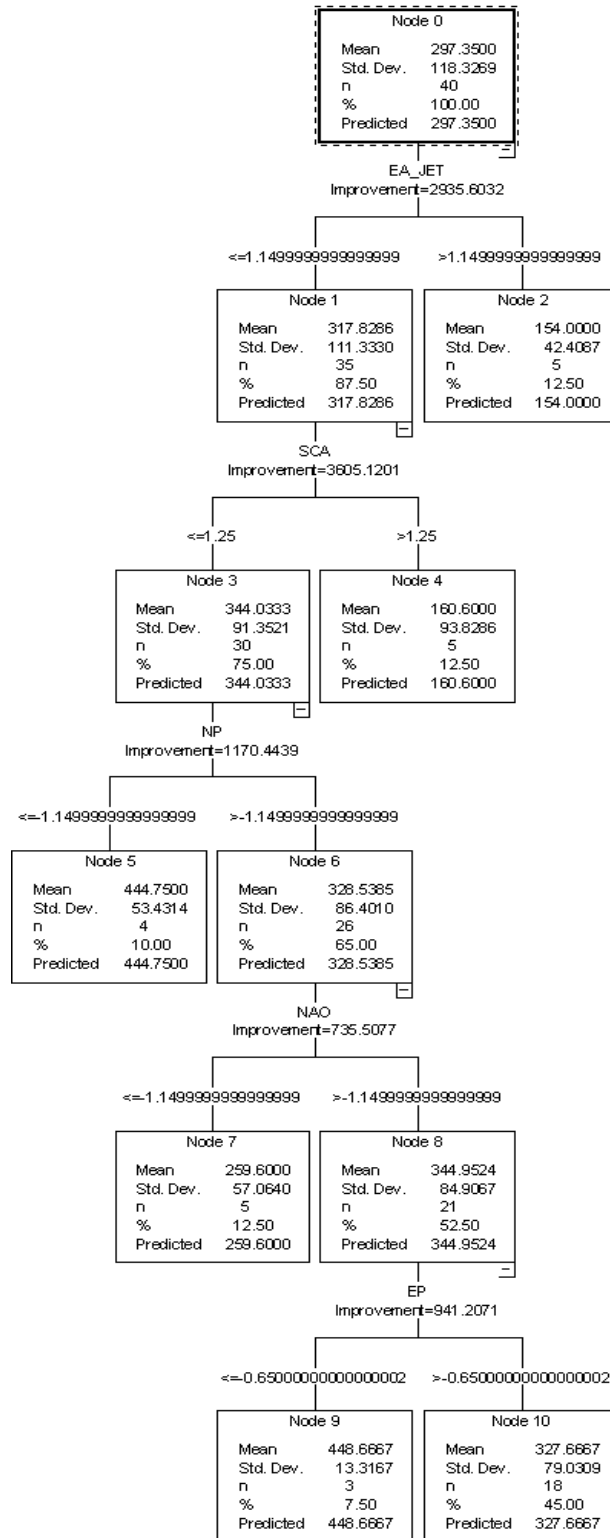


Figure 28. Portland regression tree.

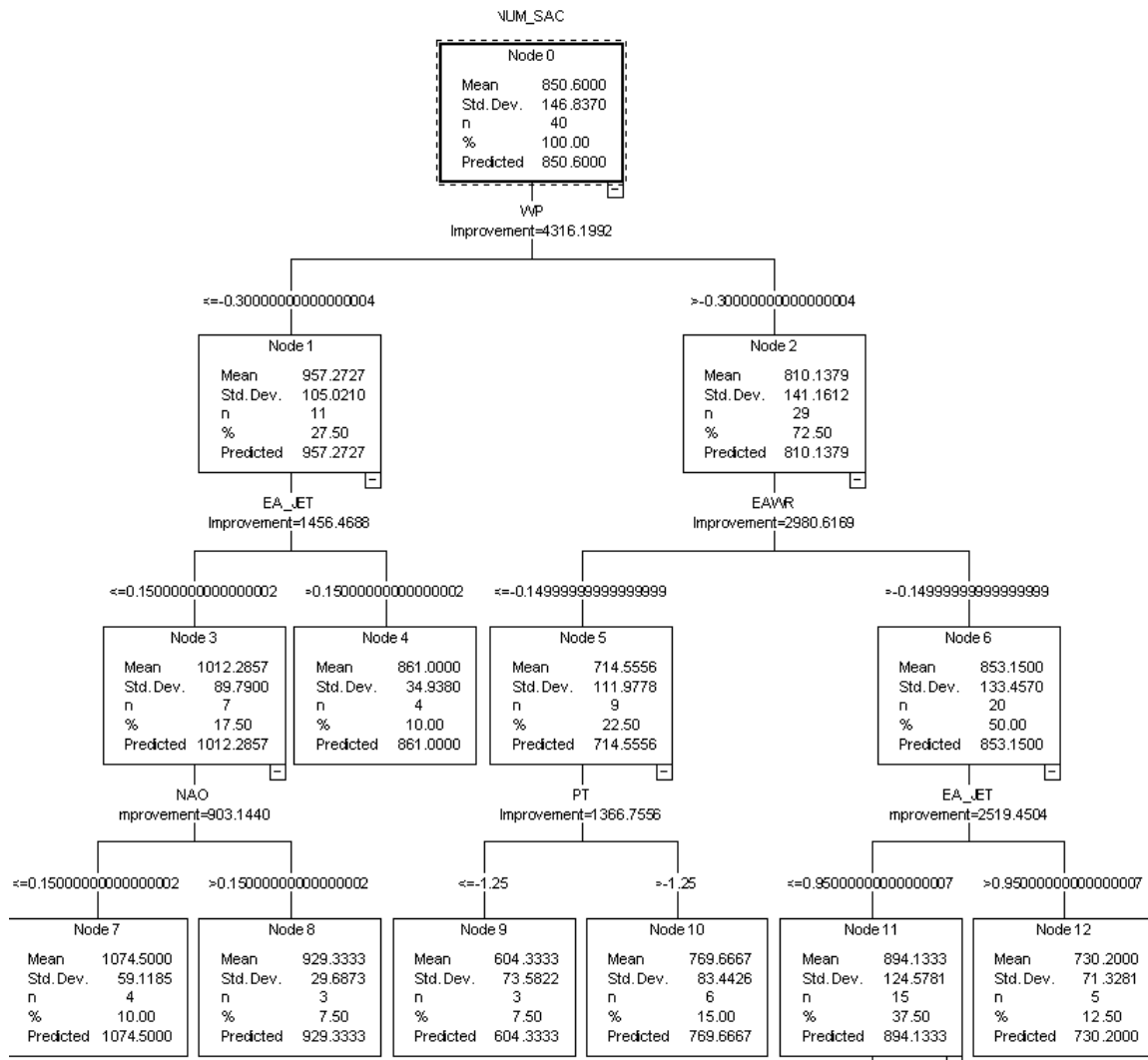


Figure 29. Sacramento regression tree.

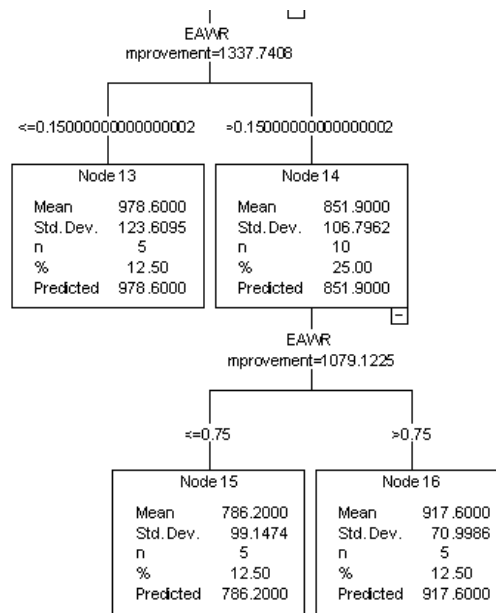


Figure 29 (continued). Sacramento regression tree.

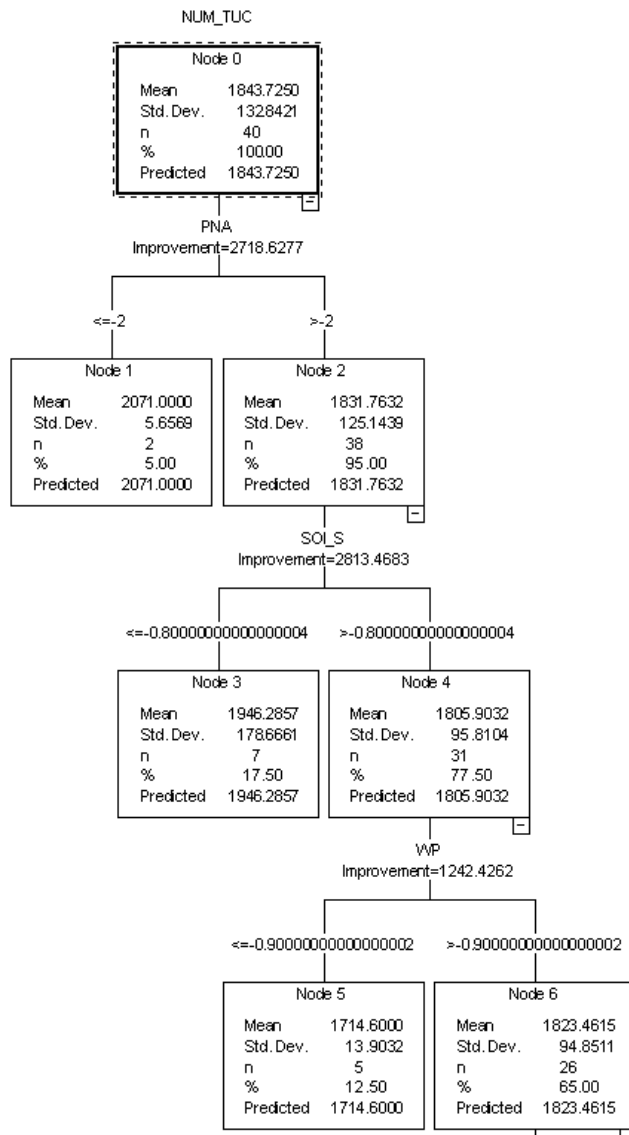


Figure 30. Tucson regression tree.

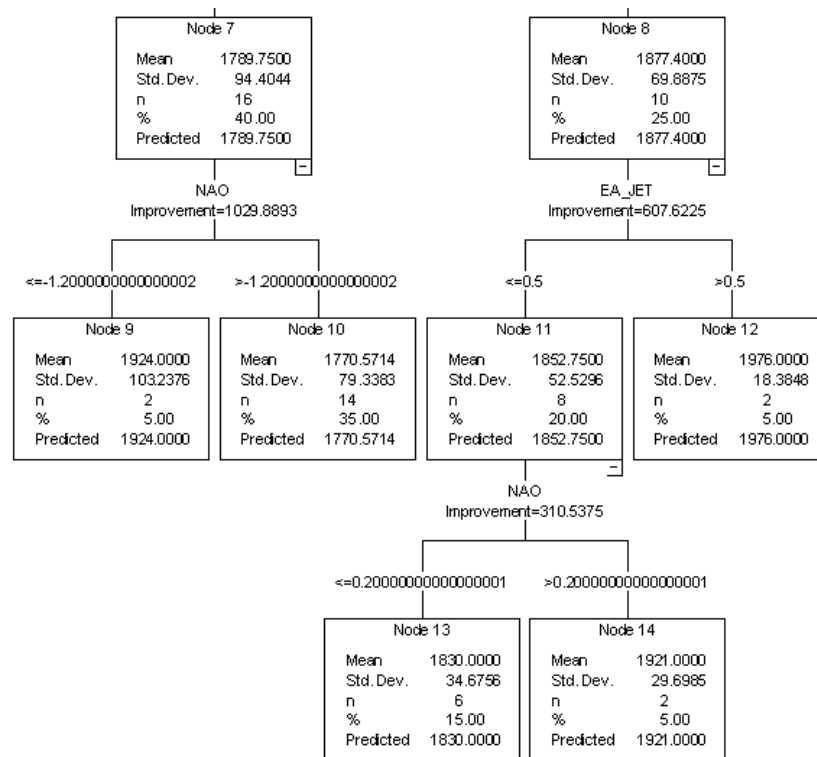


Figure 30 (continued). Tucson regression tree.



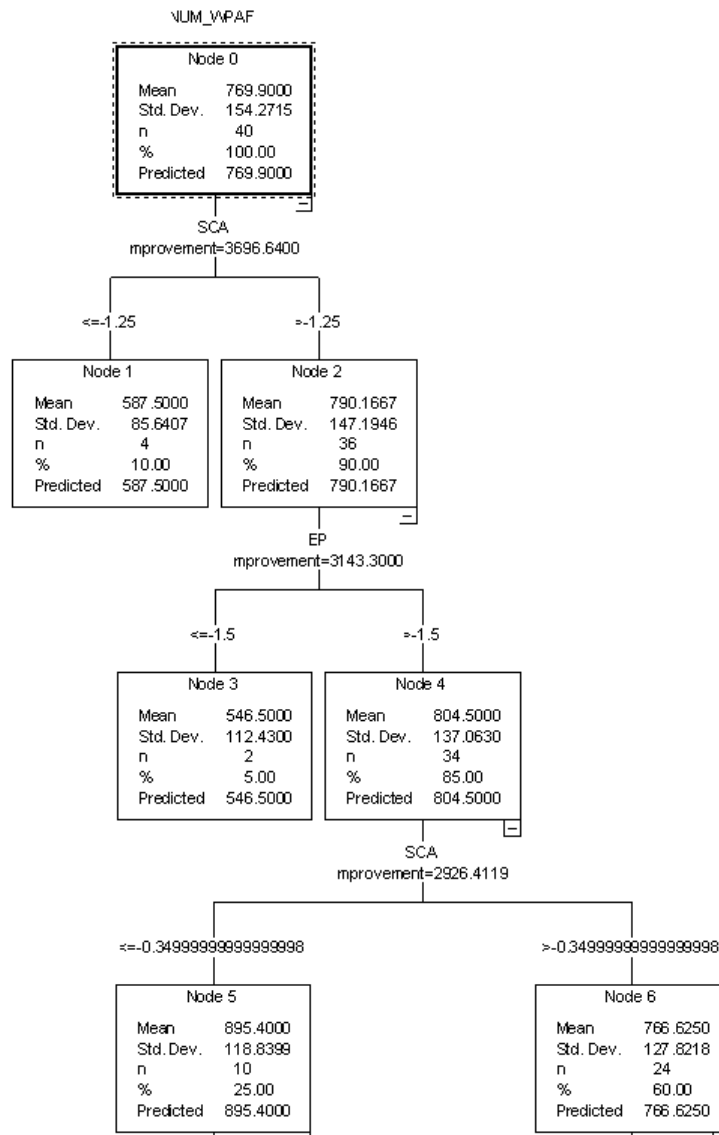


Figure 31. WPAFB regression tree.

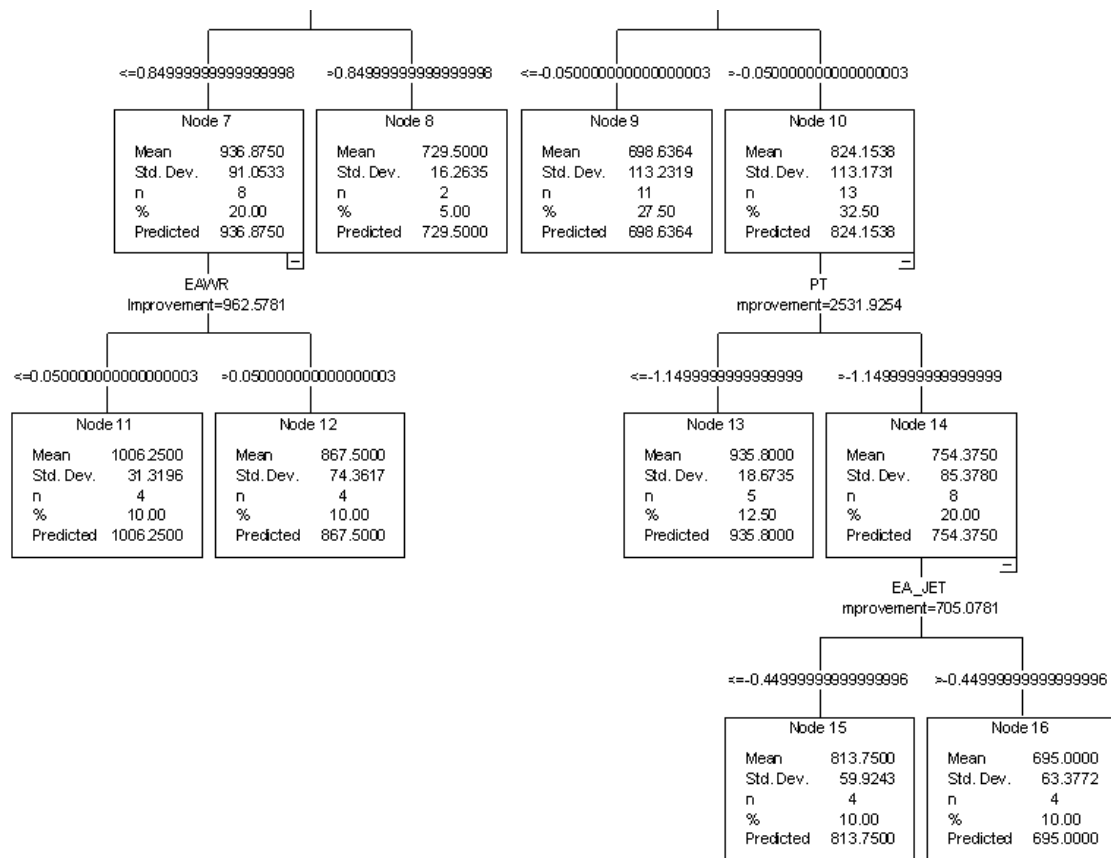


Figure 31 (continued). WPAFB regression tree.

REPORT DOCUMENTATION PAGE				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>					
1. REPORT DATE (DD-MM-YYYY) 11-03-2002		2. REPORT TYPE Master's Thesis		3. DATES COVERED (From – To) Jun 2001 – Mar 2002	
4. TITLE AND SUBTITLE  EXPLORATION OF TELECONNECTION INDICES FOR LONG-RANGE SEASONAL TEMPERATURE FORECASTS				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S)  Randall, Robb M., Captain, USAF				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S) Air Force Institute of Technology Graduate School of Engineering and Management (AFIT/EN) 2950 P Street, Building 640 WPAFB OH 45433-7765				8. PERFORMING ORGANIZATION REPORT NUMBER  AFIT/GM/ENP/02M-08	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) AFCCC/DOO Attn: Mr. Ken Walters 151 Patton Ave Rm 120 Asheville, NC 28801-5002 DSN: 673-9024 e-mail: ken.walters@afccc.af.mil				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT  APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT <p>The Air Force Combat Climatology Center (AFCCC) is tasked to provide long-range seasonal forecasts for worldwide locations. Currently, the best long-range temperature forecasts the weather community has are the climatological standard normals. This study creates a stepping-stone into the solution of long-range forecasting by finding a process to predict temperatures better than those using climatological standard normals or simple frequency distributions of occurrences. Northern Hemispheric teleconnection indices and the standardized Southern Oscillation index are statistically compared to three-month summed Heating Degree Days (HDDs) and Cooling Degree Days (CDDs) at 14 U.S. locations. First, linear regression was accomplished. The results showed numerous valid models, however, the percent of variance resolved by the models was rarely over 30%. The HDDs and CDDs were then analyzed with Data-mining classification tree statistics, however, the results proved difficult to extract any predictive quantitative information. Finally a Data-mining regression tree analysis was performed. At each conditional outcome, a range of HDDs/CDDs is produced using the predicted standard deviations about the mean. Verification of independent teleconnection indices was used as predictors in the conditional model; 90% of the resulting HDDs/CDDs fell into the calculated range. An overall average reduction in the forecast range was 35.7% over climatology.</p>					
15. SUBJECT TERMS Climatology, teleconnection indices, data mining, Classification and Regression Trees (CART), Heating Degree Days (HDD), Cooling Degree Days (CDD)					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Ronald P. Lowther, Lt Col, USAF (ENP)
U	U	U	UU	87	19b. TELEPHONE NUMBER (Include area code) (937) 255-3636, ext 4645; e-mail: ronald.lowther@afit.edu