Time series analysis of precipitation data from the Beartooth Plateau
by Deborah Kay Carlson

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Statistics
Montana State University

Abstract:
Dendroclimatologists rely on recorded weather data to correlate tree rings with historical climate patterns. Trees were cored near the location of a high elevation remote telemetry weather station so that the precipitation data recorded there could be used in a tree-ring analysis. Precipitation data from this site is compared to a more extensive record from a long-term weather station at a lower elevation. The two precipitation series are analyzed individually and together in order to compare their periodic behavior. Methods are described for estimating missing data in each of the series. Exploratory data analysis and ARIMA procedures are used to examine the behavior of the two series individually, and the two are compared in the frequency domain using bivariate analysis.

Exploratory analysis identifies several potential cycles in each data set, the most noticeable being a yearly cycle. Univariate time domain models for the shorter series include no yearly component. Models for the longer series include yearly terms as the averaging period increases. Frequency domain analysis shows that both series operate with the same set of underlying frequencies. The frequency that corresponds to a cycle with a year-long period accounts for the most variation in the data from each location.

Although time domain models drastically differ between data sets, the overall cyclic behavior of precipitation is similar. When considering the shorter series as a basis for a tree-ring model, the longer series provides a valid substitution.

TIME SERIES ANALYSIS OF PRECIPITATION DATA FROM THE BEARTOOTH PLATEAU

by

Deborah Kay Carlson

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

of

Master of Science

in

Statistics

MONTANA STATE UNIVERSITY
Bozeman, Montana

April 1995

APPROVAL

of a thesis submitted by

Deborah K. Carlson

This thesis has been read by each member of the thesis committee and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the College of Graduate Studies.

4/5/95
Date

William F. Quimby
Chairperson, Graduate Committee

Approved for the Major Department

4/5/95
Date

John Lund
Head, Statistics

Approved for the College of Graduate Studies

4/23/95
Date

Robert Brown
Graduate Dean

STATEMENT OF PERMISSION TO USE

In presenting this thesis in partial fulfillment for a master's degree at Montana State University, I agree that the Library shall make it available to borrowers under rules of the Library.

If I have indicated my intention to copyright this thesis by including a copyright notice page, copying is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for permission for extended quotation from or reproduction of this thesis in whole or in parts may be granted only by the copyright holder.

Signature Deborah K Cahn

Date 4/5/95

ACKNOWLEDGEMENTS

I would like to thank the following people without whose help and support this thesis would never have been possible.

Tim Carlson

Bill Quimby

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Dendroclimatologists rely on recorded weather data to correlate tree rings with historical climate patterns. Trees were cored near the location of a high elevation remote telemetry weather station so that the precipitation data recorded there could be used in a tree–ring analysis. Precipitation data from this site is compared to a more extensive record from a long–term weather station at a lower elevation. The two precipitation series are analyzed individually and together in order to compare their periodic behavior. Methods are described for estimating missing data in each of the series. Exploratory data analysis and ARIMA procedures are used to examine the behavior of the two series individually, and the two are compared in the frequency domain using bivariate analysis.

Exploratory analysis identifies several potential cycles in each data set, the most noticeable being a yearly cycle. Univariate time domain models for the shorter series include no yearly component. Models for the longer series include yearly terms as the averaging period increases. Frequency domain analysis shows that both series operate with the same set of underlying frequencies. The frequency that corresponds to a cycle with a year–long period accounts for the most variation in the data from each location.

Although time domain models drastically differ between data sets, the overall cyclic behavior of precipitation is similar. When considering the shorter series as a basis for a tree–ring model, the longer series provides a valid substitution.

CHAPTER 1

INTRODUCTION

An on–going research project being conducted by the Mountain Research Center(MRC), located at Montana State University– Bozeman, in Bozeman, Montana, involves reconstructing the past weather patterns in a mountainous environment using dendroclimatology, "broadly defined to include tree–ring studies involving climate–related problems [6]."The MRC has gathered tree rings from a location on the Beartooth Plateau, in northwestern Wyoming, and plans to use these rings to construct a record of past weather patterns.

In order to use tree rings to describe past weather, a model must be constructed describing the relationship between available weather data and corresponding tree rings. Nine years of weather data have been gathered via a remote telemetry weather station (SNOTEL) at the tree–coring site. The weather record contains reliable data from October, 1984 through January, 1992.

The SNOTEL telemetry system consists of several field sites located in rugged, remote locations and a main SNOTEL minicomputer in Portland, Oregon. The minicomputer "polls" SNOTEL sites, which in turn send weather information to the minicomputer by reflecting radio signals off ionized meteor trails in the upper atmosphere. This way the USDA, Soil Conservation Service(SCS) may obtain weather data, such as snow water equivalent (swe), air temperature, and soil temperature, and keep it in a database for its own uses and outside research. The data was gathered from the Beartooth Lake SNOTEL site, located 4447N, 10934W at an elevation of 9275 feet.

Because nine years represents a short period of time in terms of weather patterns and events, a longer data set is preferable. A twenty-three year weather data

set is available from a National Oceanic and Atmospheric Administration (NOAA) weather site at Cooke City, Montana. This site is northwest and at a lower elevation than the SNOTEL site. Available data begins with January, 1969 and continues through December, 1991.

The NOAA site at Cooke City is similar to many NOAA weather stations located at many airports and other strategic weather sites. The site monitors precipitation, wind and pressure variables, and some air quality variables. Its exact location is 4501N, 10958W, at an elevation of 7460 feet.

This paper has two goals. The first is to examine the relationships among the data from each site individually. Exploratory data analysis leads to familiarity with the individual data sets and helps identify interesting relationships. These relationships get further study in time and frequency domain analyses. The series are modeled in the time domain using ARIMA procedures and important cycles in the data are identified through frequency domain analysis.

The second goal is to describe the relationship between the SNOTEL site data and the NOAA site data to determine whether the longer NOAA data set may be used as a basis for the tree–ring model. Once the individual analyses are complete, the two series are considered simultaneously in a bivariate analysis. Exploratory data analysis helps identify important relationships between the two which are further examined in the frequency domain.

CHAPTER 2

TIME DOMAIN ANALYSIS

Theory of Stationary Random Processes

A model is a set of assumptions about the mathematical process that may have generated some observed data [7]. One goal of time series analysis is to build a model that concisely describes the process that generated an observed series. The nature and complexity of time series often requires the model to be a function of time. A commonly used model building method follows the Autoregressive Integrated Moving Average(ARIMA) modeling approach. It is based on the idea that a time series is a single realization from an infinite number of possibilities that may be generated by the underlying process. For example, the amount of precipitation that occurs on a given day is one realization of the amount that could have occurred. If time were able to repeat itself, a different amount of precipitation could occur that day. If a day were repeated over and over, the rainfall that occurs would form a distribution(pdf). The same holds true for every day of a time series–the data form a single realization from a pdf. Thus, a time series is a collection of random variables whose pdf's describe something about the structure of the underlying process. This collection of random variables is referred to as a random process. All possible realizations of a random process form the ensemble.

Past events provide little insight into the current state of an unstable or completely random fluctuating process. An implicit assumption when using time series analysis to describe an underlying process is that it is in some way stable. Scientists

assume the underlying process driving precipitation is stable and certain precipitation patterns continually repeat themselves.

One form of the stability conditions mentioned above is called second order stationarity(SOS) conditions. Identifying these conditions requires notation that statistically describes a times series. A time series denoted $\{x_t,\ t = 1, 2, \ldots, T\}$ is a realization of the random process $\{X_t,\ t = 1, 2, \ldots, T\}$. The mean of X_t at time t is $\mathrm{E}[X_t] = \mu_t$. The autocovariance between X_t and X_{t+k} is

$$\gamma(t, t+k) = \mathrm{cov}(X_t, X_{t+k}) = \mathrm{E}[(X_t - \mu_t)(X_{t+k} - \mu_{t+k})] \tag{2.1}$$

where $k = 0, 1, 2, \ldots, T-1$ is called the lag.

Second order stationarity is defined as follows: $\{X_t,\ t = 1, 2, \ldots, T\}$ is an SOS process if it has a constant mean and variance over time and the autocovariance between two time locations depends only on the lag between them. When these conditions are met, $\gamma(0)$ represents the variance σ^2. These conditions are summarized as follows:

1. $\mu_t = \mu\ \forall\ t$
2. $\sigma_t^2 = \sigma^2\ \forall\ t$
3. $\gamma(t, t+k) = \gamma(k)\ \forall\ t.$

From conditions 2 and 3, the autocorrelation function may be defined as

$$\rho(k) = \frac{\gamma(k)}{\gamma(0)}. \tag{2.2}$$

It exhibits the following properties:

1. $\rho(k) = \rho(-k)$
2. $-1 \leq \rho(k) \leq 1$

3. $\rho(k) = 0$ if X_t and X_{t+k} are independent.

A stationary random process is ergodic if the ensemble average (of means, variances, and covariances) at any point in time equals the corresponding time average [7]. Ergodicity implies that a single realization contains as much information about the mean and covariance structure of the random process as ensemble means and covariances [10]. It allows the formulation of a model for the random process through a single observed time series. The autocorrelation function provides the basic characterization of a stationary random process.

The k^{th} sample autocovariance coefficient takes the form:

$$g_k = \frac{1}{T} \sum_{t=k+1}^{T} (x_t - \bar{x})(x_{t-k} - \bar{x}), \tag{2.3}$$

where $\bar{x} = \frac{1}{T} \sum_{t=1}^{T} x_t$ and g_k estimates $\gamma(k)$. It follows that the k^{th} autocorrelation coefficient is then

$$r_k = \frac{g_k}{g_0}, \tag{2.4}$$

where g_0 is the sample variance; r_k estimates ρ_k. The correlogram of $\{x_t\}$ is the graph of k vs. r_k.

Analysts often examine the correlogram to determine whether a time series is white noise. Bartlett showed that $r_k \sim N\left(0, \frac{1}{T}\right)$ for large T if x_t is a white noise sequence [5]. Using this result, 95% significance bands are drawn on the correlogram at $\pm\frac{2}{\sqrt{T}}$. Any r_k falling outside these bands could be significant at the .05 level. When r_k is calculated for many k, some r_k may fall outside the significance bands and not be significant due to sampling variability. However, the significance bands provide an initial indication of a white noise sequence.

Box & Pierce developed the portmanteau test–statistic for white noise [3]. Ljung & Box refined it in 1978. They showed that for a white noise sequence with

large T and v much smaller than T,

$$Q_v = T(T+2)\sum_{i=1}^{v}\frac{1}{T-i}r_i^2 \sim \chi^2(v). \tag{2.5}$$

Q_v is the test–statistic for $H_o : \rho_1 = \rho_2 = \ldots = \rho_v = 0$ versus H_A : at least one differs. H_o implies white noise under the assumption of an independent, identically distributed Gaussian error structure.

ARIMA Modeling Procedures

Once stationarity and a dependent structure within the data have been established, the correlogram also helps describe the nature of the dependence between observations. Comparing the overall shape of the correlogram to theoretical autocorrelation functions of certain models establishes potential models for an observed time series.

Using ARIMA methods, plausible models take the form of autoregressive integrated moving average (ARIMA) models. They are combinations of simpler autoregressive, integrated, and moving average components. Special notation has been established to identify the separate components contributing to the model: p and q specify the orders of the autoregressive and moving–average components, respectively, while d represents the level of integration. The order of the chosen model is specified using the notation ARIMA(p, d, q).

Special cases of ARIMA(p, d, q) notation occur when the chosen model includes only one type of component. An ARIMA$(p, 0, 0)$ model is an autoregressive model of order p, denoted AR(p). X_t is modeled as a linear combination of p previous X's and white noise. The $AR(p)$ model is expressed as

$$\begin{aligned} X_t &= a_0 + a_1X_{t-1} + a_2X_{t-2} + \cdots + a_pX_{t-p} + \varepsilon_t \\ \Rightarrow X_t &= a_0 + \sum_{i=1}^{p} a_{t-i}X_i + \varepsilon_t \end{aligned} \tag{2.6}$$

where $\varepsilon_t \sim iid(0, \sigma^2)$ and $a_j,\ j = 0, 1, 2, \ldots, p$ are parameters.

An ARIMA$(0, 0, q)$ model denotes a moving average model of order q, denoted MA(q). X_t is modeled as a linear combination of q previous error terms and white noise¿ The $MA(q)$ model is expressed as

$$\begin{aligned} X_t &= \varepsilon_t - b_1\varepsilon_{t-1} - b_2\varepsilon_{t-2} - \cdots - b_q\varepsilon_{t-q} - b_0 \\ \Rightarrow X_t &= \varepsilon_t - b_0 - \sum_{k=1}^{q} b_{t-k}\varepsilon_{t-k} \end{aligned} \tag{2.7}$$

where $\varepsilon_t \sim iid\,(0, \sigma^2)\,,\ t = 1, 2, \ldots, T$ and $b_l,\ l = 0, 1, \ldots, q$ are parameters.

An ARIMA$(0, d, 0)$ is an integrated model with d differences required to make the sequence stationary. A classical example of an integrated random sequence is a random walk. The variable of interest is a sum of *iid* random variables with outcomes -1 or 1, each of which is equally likely. In general, an ARIMA$(0, d, 0)$ model says that the time series has been reduced to a white noise sequence through differencing d times [9].

Most time series without a seasonal component may be modeled as a combination of AR(p), MA(q), and integrated models. Modeling involves three steps to identify the final ARIMA(p, d, q) form: identification of the order(determining p, d, and q); estimation of the model parameters(the a's and b's); and diagnostic checking of the candidate model fit. If diagnostic checks find the candidate model lacking, the three step procedure begins again with a new candidate model.

Identification begins with an examination of the series for stationarity. If the series appears stationary, then $d = 0$. Some authors describe differencing and other methods of handling nonstationarity [9], [7], [5]. Once stationarity is achieved, the next step in identification involves examining the autocorrelation structure of the data. Because two or more series with qualitative differences may share the same correlogram, a partial correlogram is also calculated. It is described below.

Without loss of generality, the mean of the series may be set to zero so that an alternate form to (2.6) of an AR(p) process is

$$X_t = \sum_{i=1}^{p} a_i X_{t-i} + \varepsilon_t. \tag{2.8}$$

Multiplying (2.8) by X_{t-k} gives

$$X_t X_{t-k} = \sum_{i=1}^{p} a_i X_{t-i} X_{t-k} + \varepsilon_t X_{t-k}$$

Applying the expectation operator gives

$$\begin{aligned} \mathrm{E}\left[X_t, X_{t-k}\right] &= \mathrm{E}\left[\sum a_i X_{t-i} X_{t-k} + \varepsilon_t X_{t-k}\right] \\ \gamma(k) &= \sum a_i \mathrm{E}\left[X_{t-i} X_{t-k}\right] + 0 \\ \Rightarrow \gamma(k) &= \sum a_i \gamma(k-i), \quad k = 1, 2, \ldots. \end{aligned} \tag{2.9}$$

For $k > 0$, X_{t-k} and ε_k are independent and $\mathrm{E}\left[\varepsilon_t\right] = 0$. Dividing both sides by γ_0 gives a system of equations in terms of the autocorrelation coefficients:

$$\rho(k) = \sum_{i=1}^{p} a_i \rho_{k-i}, \ k = 1, 2, \ldots. \tag{2.10}$$

Thus, the autocorrelation coefficients may be expressed as functions of the a's. This system of equations is known as the Yule–Walker equations [5]. The sample version of these equations is

$$r_k = \sum_{i=1}^{p} \hat{a}_i r_{k-i}, \ i = 1, 2, \ldots, p, \tag{2.11}$$

where $\hat{a}_i, \ i = 1, 2, \ldots, p$ estimate $a_i, \ i = 1, 2, \ldots, p$ for an AR(p) process. Writing this system in matrix notation gives

$$\vec{r} = R\vec{\hat{a}},$$

where the $(i, j)^{th}$ element if $R_{p \times p}$ is r_{i-j}. Inverting R gives,

$$\vec{\hat{a}} = R^{-1}\vec{r}.$$

The p^{th} partial autocorrelation estimate is $\hat{a}_p$. p partial autocorrelation coefficient estimates are calculated by successively fitting AR(1), AR(2), ..., AR(p) processes to the data, getting a partial autocorrelation coefficient estimate with each fit. A partial autocorrelation function(pacf) is then a graph of k versus r_k. The partial correlogram estimates the pacf.

Examination of the correlogram and partial correlogram determines whether the series is a realization from an AR or MA random process and provides initial values for p or q. Quenouille(1949) showed that, for an underlying AR(p) random process, $a_k \sim N(0, \frac{1}{T})$ for $k > p$. Thus, .05 significance bands drawn on the partial correlogram at $\pm\frac{2}{\sqrt{T}}$ provide an initial indication of the order of the process. If the underlying process is an AR(p), the partial correlogram "cuts off" to zero at lag $k = p$ and the correlogram "tails off" slowly. A "cut off" occurs when a spike at lag k is significant and the spike at lags $k+1, k+2, \ldots$ are insignificant and much closer to zero than the value at lag k. A "tail off" occurs when values of the function start out significant and slowly decrease in significance as the lag increases. When the underlying process is an MA(q) the correlogram cuts off at $k = q$ and the partial correlogram "tails off." No cut off in either the correlogram or partial correlogram indicates a mixed ARMA model.

If the correlogram and partial correlogram indicate an underlying AR(p) random sequence, the Yule–Walker equations are used to estimate the coefficients. When an AR(p) random sequence is written in the form of equation (2.8) the coefficient estimates are found by solving the sample version of the Yule–Walker equations of the form (2.11).

Identification and coefficient estimation procedures are straightforward for series without seasonal autocorrelations. However, a seasonal component is difficult to incorporate in the model using sequential ARIMA fitting procedures. In order

to include a seasonal component in a monthly series, the model would require 12 significant coefficients: a_1 through a_{12}.

Many scientists avoid the seasonal component of a time series by differencing the series in the appropriate fashion enough times to remove it. For example, if a time series $\{y_t,\ t = 1, 2, \ldots, T\}$ contains monthly observations, then the differenced time series $\{z_i,\ i = 1, 2, \ldots, n = T - 12\}$ would consist of the observations

$$\begin{aligned} z_1 &= (y_{13} - y_1) \\ z_2 &= (y_{14} - y_2) \\ &\vdots \\ z_n &= (y_T - y_{T-12}). \end{aligned}$$

If this series still exhibits seasonal fluctuation, it is differenced accordingly until the seasonal component is completely removed.

Differencing removes the seasonal component rather than incorporating it into a model. Because the seasonal component is such an integral part of many natural phenomena, the ARIMA$(p, d, q)(P, D, Q)_S$ method was developed to include it as a model component. The ARIMA(p, d, q) method uses the regular (p, d, q) portion to describe the within–season dependence and the $(P, D, Q)_S$ portion to describe the seasonal dependence [9]. S specifies the number of observations in a period. P and Q specify the orders of the seasonal autoregressive and moving–average components, and D represents the number of seasonal differences required for stationarity.

As an example, consider a time series composed of monthly observations believed to follow a yearly cycle. Then $S = 12$. If observations 12 time units apart exhibit a trend, the series is differenced as described above: $z_1 = (y_{13} - y_1), z_2 = (y_{14} - y_2)$, etc., until the trend is removed. D specifies the required number of differences to remove the seasonal trend. Once the series is seasonally stationary, the order of seasonal autoregressive and moving average components can be determined.

In general, the regular and seasonal components of an ARIMA$(p, d, q)(P, D, Q)_S$ model are of the same type [9]. In other words, if the regular process is an AR(p) ran-

dom process and a seasonal component exists, it is usually autoregressive. The same holds true for moving average models. In practice if a series exhibits nonstationarity, many times regular **or** seasonal differencing corrects the problem.

The identification of P and Q follows the same procedure as determining p and q, only with the seasonal lags. In general, the seasonal lags are denoted S, $2S$, $3S$, etc. Purely seasonal random processes take one or a combination of the following forms:

$$\text{AR}(P): \qquad X_t = \sum_{j=1}^{P} \tau_j X_{jS} + \varepsilon_t \tag{2.12}$$

$$\text{MA}(Q): \qquad X_t = \varepsilon_t - \sum_{j=1}^{Q} \theta_j \varepsilon_{jS}. \tag{2.13}$$

The correlogram at lags S, $2S$, ... tails off to zero and the same lags of the partial correlogram cut off to zero at PS for a realization from an AR(P) seasonal random process. When the correlogram cuts off to zero at lag QS and seasonal lags of the partial correlogram tail off, an MA(Q) model may be appropriate. Large and nearly equal values of the correlogram at the seasonal lags indicates seasonal nonstationarity [9]. A mixed seasonal model is appropriate if neither the correlogram nor partial correlogram tail off at seasonal lags.

After selecting a candidate model and estimating the coefficients, the analyst uses residual diagnostics to assess the quality of fit as an indicator of model adequacy. Two methods of determining model adequacy include checking independence of residuals at the first two regular and seasonal lags and comparing the residual series to a white noise sequence. Observations can be written as the sum of the fit component and the residual component; for an AR(p)(P)$_S$ sequence,

$$\begin{aligned} x_t &= \sum_{i=1}^{p} \hat{a}_i x_{t-i} + \sum_{j=1}^{P} \hat{\tau}_j x_{t-jS} + e_t \\ \Rightarrow e_t &= x_t - \sum_{i=1}^{p} \hat{a}_i x_{t-i} - \sum_{j=1}^{P} \hat{\tau}_j x_{t-jS} \end{aligned} \tag{2.14}$$

where $\{e_t,\ t = 1, 2, \ldots, t\}$ is the residual series and $\hat{\tau}_j$ estimates the j^{th} seasonal coefficient.

If the residuals are independent at the first two regular lags, i.e. $\mathrm{E}[e_t, e_{t+1}] = \mathrm{E}[e_t, e_{t+2}] = 0$, then the correlogram at $k = 1, 2$ of the residual series equals zero: $r_1 = r_2 = 0$. The seasonal lags must follow the same criterion. In order for $\mathrm{E}[e_t, e_{t+S}] = \mathrm{E}[e_t, e_{t+2S}] = 0$, partial correlogram values at r_S and r_{2S} must be zero. The regular correlogram error bars $\pm\frac{2}{\sqrt{T}}$ may be applied to the residual correlogram with a note of caution: the standard error of the series correlogram may underestimate the standard error of the residual correlogram [2]. Therefore, one should consider the error bars an upper limit for significant values of the residual correlogram.

If the residual series is a white noise sequence, the residual correlogram and partial correlogram are uniformly zero for all lags. A qualitative test involves graphing the residual correlogram and partial correlogram with error bars, keeping in mind the cautionary note above. The residual series may not be white noise if many ordinates exceed the error bars. The portmanteau test–statistic in equation (2.5) can be used here to test for correlations among the residuals. A suitable value for v corresponds to two years worth of data to test for seasonal autocorrelations [9].

A quantitative diagnostic called "trial overfitting" involves increasing p, P, or both by one and assessing whether the additional parameter(s) significantly improves the fit. The residual mean square(RMS) of the candidate and overfit models are compared, where

$$RMS = \frac{1}{T}\sqrt{\sum_{i=1}^{T} e_t^2}. \tag{2.15}$$

A significant reduction in RMS from the candidate model to the overfit signals potential inadequacy of the candidate. Another quantitative measure of model adequacy involves spectral checking which is discussed in the frequency domain analysis section.

CHAPTER 3

SPECTRAL ANALYSIS

Univariate Theory

One may reasonably assume that weather variables follow patterns that are repeated year after year. Certain weather variables, such as precipitation, might also follow a pattern within a season that may repeat itself four times a year. In fact, precipitation may consist of several cycles undetected in exploratory analysis or ARIMA modeling. Spectral analysis identifies the frequencies that account for a significant portion of the variation in a time series. It uses Fourier transforms to divide the series into its major frequency components by estimating the amplitudes of the Fourier frequencies. Those with significant amplitudes are considered the component frequencies of the series. Spectral analysis assumes that a time series $\{y_t, \quad t = 1, 2, \ldots, T\}$ may be written as the sum of an infinite number of sinusoids cycling at different frequencies:

$$y_t = \sum_{j=0}^{\infty} R_j \cos(\omega_j t + \phi_j), \quad t = 1, 2, \ldots, T. \tag{3.1}$$

The model used to describe this infinite sum identifies $\{y_t\}$ as the sum of m Fourier frequencies and white noise so that

$$y_t = \sum_{j=1}^{m} R_j \cos(\alpha_j t + \phi_j) + \varepsilon_j, \quad t = 1, 2, \ldots, T[1], \tag{3.2}$$

where $\alpha_j = \frac{2\pi j}{T}$ is the j^{th} Fourier frequency and $\varepsilon_j \sim iidN(0, \sigma^2)$. R_j and ϕ_j are parameters representing the amplitude and phase associated with α_j. Here m is the

largest integer less than $\frac{T}{2}$. $\frac{T}{2}$ is the Nyquist frequency, the smallest frequency detectable using the discrete Fourier transform of (3.2). The Nyquist frequency repeats itself every other time unit, so that its period is two.

Currently, (3.2) is not linear in either R or ϕ. However, a linear models framework may be applied to the model if it is rewritten to be linear in the parameters. Using the identity $\cos(a+b) = \cos a \cos b - \sin a \sin b$, the linear model becomes

$$y_t = \sum_{j=1}^{m} A_j \cos(\alpha_j t) + B_j \sin(\alpha_j t) + \varepsilon_j, \quad j = 1, 2, \ldots, T, \tag{3.3}$$

where $A_j = R_j \cos(\phi_j)$, $B_j = -R_j \sin(\phi_j)$, and $\varepsilon_j \sim iidN(0, \sigma^2)$[5].

The least–squares estimates of A_j and B_j are

$$\hat{A}_j = \frac{2}{T} \sum_{t=1}^{T} y_t \sin(\alpha_j t), \tag{3.4}$$

$$\hat{B}_j = \frac{2}{T} \sum_{t=1}^{T} y_t \cos(\alpha_j t). \tag{3.5}$$

Estimates of R_j and ϕ_j are then

$$\hat{R}^2 = \hat{A}^2 + \hat{B}^2$$

and

$$\hat{\phi} = \begin{cases} \arctan(-\hat{B}/\hat{A}), & \hat{A} > 0 \\ \arctan(-\hat{B}/\hat{A}) - \pi, & \hat{A} < 0, \hat{B} > 0 \\ \arctan(-\hat{B}/\hat{A}) + \pi, & \hat{A} < 0, \hat{B} \leq 0 \\ -\pi/2, & \hat{A} = 0, \hat{B} > 0 \\ \pi/2 & \hat{A} = 0, \hat{B} < 0 \\ \text{arbitrary} & \hat{A} = 0, \hat{B} = 0. \end{cases} \quad [1]$$

The "strength" of frequency ω often gets displayed by a plot of the spectral density function. It is a transformation of the autocovariance function γ_k. and is written

$$\begin{aligned} f(\omega) &= \sum_{k=-\infty}^{\infty} \gamma_k \exp(-i\omega k) \\ &= \gamma_0 + 2\sum_{k=1}^{\infty} \gamma_k \cos(\omega k). \end{aligned}$$

A graph of $f(\omega)$ versus ω exhibits a peak at every component frequency.

After transforming $\{y_t\}$ via the discrete Fourier transform, only values of the spectral density function at the Fourier frequencies, $\alpha_j, \quad j = 1, 2, \ldots, m$ are considered. The periodogram, denoted $I(\alpha)$, estimates the spectral density function, $f(\alpha)$. The j^{th} periodogram ordinate is written

$$\begin{aligned} I(\alpha_j) &= \frac{1}{T}\left[\left(\sum_t y_t \sin(\alpha_j t)\right)^2 + \left(\sum_t y_t \sin(\alpha_j t)\right)^2\right] \\ &= \frac{T}{4}[\hat{A}^2 + \hat{B}^2]. \end{aligned}$$

The periodogram has a $\frac{\sigma^2\chi^2(2)}{2}$ distribution, resulting in a mean and variance of σ^2 and σ^4, respectively. Consequently, $I(\alpha)$ is an unbiased but inconsistent estimator of $f(\alpha)$. The most common technique for correcting the inconsistency factor involves smoothing the periodogram with a Daniell window [7]. The Daniell window applies a discrete spectral average of order $2p+1$ to $I(\alpha)$ which results in a fitted value of

$$\hat{f}(\alpha_j) = \frac{1}{2p+1}\sum_{t=-p}^{p} I(\alpha_{j+t}).$$

The independence of the Fourier frequencies implies that both the periodogram and the smoothed periodogram ordinates are independent [7]. Therefore, the smoothed periodogram may be considered frequency by frequency or within bands of frequencies when searching for those that account for a significant amount of variation in the data.

A form of the periodogram also serves as a diagnostic check for a candidate ARIMA model. The residuals from an adequate ARIMA model form a white noise sequence. All frequencies contribute equally to a white noise sequence, so the spectral density function is zero over the set of Fourier frequencies. Smoothed periodogram ordinates, when graphed cumulatively, should increase from zero to one approximately linearly [5]. To test this, let $C_j = \sum_{k=1}^{j} I(\alpha_j) : \qquad j = 1, 2, \ldots, m$. Then the j^{th} cumulative periodogram ordinate is defined as $E_j = C_j/C_m$ and the cumulative periodogram

is a graph of E_j versus j/m^*, $m^* = m - 1$[5]. Tolerance bands plotted on the cumulative periodogram help measure the departure from linearity. The tolerance bands are based on the Kolmogorov–Smirnov statistic that tests whether data, in this case the E_j, are an ordered random sample from a $Unif(0,1)$ distribution [5].

Bivariate Theory

When comparing two time series in the frequency domain, the series must be the same length so that they share the same set of Fourier frequencies. Independence of the Fourier frequencies allow for comparison of the two series frequency by frequency via their phases and amplitudes. For a given frequency, the cross–amplitude measures the covariance between the amplitudes from each series and the relative phase measures the extent one series leads the other. The relationship between two time series sharing the same components can be described by a comparison of the amplitudes and phases of the corresponding frequencies.

For a specified frequency, α, the components of the stationary bivariate process $\{X_t, Y_t\}$ can be written from equation (3.2),

$$x_t = R_x \cos(\alpha t + \phi_x),$$
$$y_t = R_y \cos(\alpha t + \phi_y).$$

The complex exponential equivalent of these equations is

$$x_t = R_x \exp(i(\alpha t + \phi_x)),$$
$$y_t = R_y \exp(i(\alpha t + \phi_y)).$$

From this, the cross–spectrum is defined as the covariance

$$\mathrm{E}[X_t, Y_t] = \mathrm{E}[R_x R_y]\mathrm{E}[\exp(i(\phi_x - \phi_y))][5].$$

The cross–amplitude spectrum, denoted $R_{xy}(\alpha)$, is a plot of $\mathrm{E}[R_x R_y]$ as a function of frequency. Similarly, the phase spectrum, denoted $\phi_{xy}(\alpha)$, is a plot of $\mathrm{E}[\exp(i(\phi_x - \phi_y))]$ as a function of frequency. Thus, the cross–spectrum can be written

$$f_{xy}(\alpha) = R_{xy}(\alpha)\exp(i\phi_{xy}(\alpha))[7].$$

Oftentimes a quantity known as the coherency is plotted rather than the cross–amplitude spectrum. The coherency at frequency α is written

$$\rho^2(\alpha) = \frac{|f_{xy}(\alpha)|^2}{f_x(\alpha)f_y(\alpha)},$$

where $f_x(\alpha)$ and $f_y(\alpha)$ are the marginal spectra of $\{X_t\}$ and $\{Y_t\}$, respectively.

The individual spectra, $f_x(\alpha)$ and $f_y(\alpha)$, measure the variances between frequency components of the individual processes $\{X_t\}$ and $\{Y_t\}$. Similarly, the cross–spectrum measures the covariance between frequency components of the individual processes for a specified frequency, α. Thus the coherence is a measure of the squared correlation between corresponding frequency components of $\{X_t\}$ and $\{Y_t\}$. Because coherence corresponds to squared correlation, the association between the marginal amplitudes is not clear. However, one can estimate the values of the marginal amplitudes from the least–squares estimates given by (3.5).

The phase spectrum measures the lead–lag relationship between corresponding components of $\{X_t\}$ and $\{Y_t\}$ for each Fourier frequency. When all frequencies share a constant phase shift, the phase spectrum will be a straight line whose slope is equal to the time lag. A positive slope indicates $\{X_t\}$ leads $\{Y_t\}$ and vice versa for a negative slope [5],[7].

The cross–periodogram estimates the cross–spectrum, just as the periodogram estimates the marginal spectrum. The j^{th} cross–periodogram ordinate is

$$I(\alpha_j) = \frac{1}{T^*}J_x(\alpha_j)J_y^*(\alpha_j), \tag{3.6}$$

where J^* is the complex conjugate of J and

$$J_x = \sum_{t=1}^{T^*} x_t \exp(-i\alpha_j t)[1].$$

The cross–periodogram is an inconsistent estimator of the cross–spectrum, just as the periodogram inconsistently estimates the spectrum. Smoothing the cross–periodogram with a Daniell window smoother corrects for the inconsistency. The j^{th} smoothed cross–periodogram ordinate is

$$\hat{f}_{xy}(\alpha_j) = \frac{1}{T^*} \sum_{t=-p}^{p})I(\alpha_{j+t}). \tag{3.7}$$

Once the periodogram is smoothed, estimates of the coherency and phase can be extracted [5].

Because coherency measures correlation, it is bounded by the interval $(0, 1)$. Likewise for its estimator r^2_{xy}, which can be written

$$r^2_{xy}(\alpha) = \frac{\hat{f}_{xy}(\alpha)}{\hat{f}_x(\alpha)\hat{f}_y(\alpha)}, \tag{3.8}$$

where $\hat{f}_x$ and $\hat{f}_y$ are the smoothed marginal periodograms of $\{x_t\}$ and $\{y_t\}$, respectively.

The estimated coherency spectrum is a plot of r as a function of frequency. The series cycle together at frequencies where both the marginal spectral estimates peak together and coherency is high. Low coherence implies little correlation between amplitudes, making a phase difference between the two unimportant. Therefore, the relative phase spectrum provides information about the phase shift only at frequencies with high coherence. Relative phase estimates at those frequencies with low coherence fluctuate randomly [7].

In order to get the statistic that estimates relative phase, it is convenient to expand (3.6) using Euler's equality: $\exp(i\alpha) = \cos(\alpha) + i\sin(\alpha)$ so that

$$\begin{aligned} I(\alpha) &= \tfrac{1}{T^*}\left[A_x(\alpha)A_y(\alpha) + B_x(\alpha)B_y(\alpha)\right] - \tfrac{1}{T^*} i\left[A_y(\alpha)B_x(\alpha) - A_x(\alpha)B_y(\alpha)\right] \\ &= \tilde{c}(\alpha) - i\tilde{q}(\alpha) \end{aligned} \tag{3.9}$$

where $A_x = \sum_{t=1}^{T^*} \cos(\alpha t)$ and $B_x = \sum_{t=1}^{T^*} \sin(\alpha t)$[8]. Here, $\tilde{c}$ and $\tilde{q}$ are estimators of quantities known as the co–spectrum and quadrature spectrum. The smoothed periodogram of (3.7) can be written

$$\hat{f}_{xy} = \hat{c} - i\hat{q} \ .$$

where $\hat{c}$ and $\hat{q}$ are smoothed using a Daniell window smoother [5].

Once the co–spectrum and quadrature spectrum are smoothed, the relative phase estimator is

$$\hat{\phi}_{xy}(\alpha) = \arctan\left(\frac{-\hat{q}(\alpha)}{\hat{c}(\alpha)}\right). \tag{3.10}$$

The relative phase spectrum measures the extent to which one series leads the other. A constant time lag appears in the phase spectrum as a straight line with a slope proportional to the time lag between the two series. Oftentimes, the ordinates of $\hat{\phi}_{xy}$ are standardized to fall in the interval $[0, 2\pi)$ so that ordinates with a value of 2π represent zero phase shift. Because of this, the graph of $\hat{\phi}_{xy}$ may have several large discontinuities where the ordinates jump from near zero to near 2π.

An approximate $C\%$ pointwise confidence interval for ϕ_{xy} is

$$\hat{\phi}_{xy} \pm \arcsin\left(t^* \left[\frac{1}{2T^* - 2}\frac{1 - r_{xy}^2(\alpha)}{r_{xy}^2(\alpha)}\right]^{1/2}\right), \tag{3.11}$$

where t^* is the upper $C\%$ value of the Student's t distribution with $(2T^* - 2)$ degrees of freedom [8]. Using the marginal spectrum and coherence estimates, one can determine which frequencies to examine in the estimated relative phase plot. Ordinates near 0 or 2π indicate no phase shift, while a phase shift of π indicates a difference of half a period.

CHAPTER 4

METHODS

Univariate Exploratory Data Analysis

Both the NOAA and SNOTEL data sets contain missing observations, extreme values, or both in the time frame common to both series. A goal of the individual series analyses is to determine whether estimating the missing data or removing the extreme values and re–estimating them will affect comparisons between the two. Therefore, the individual analyses involve comparisons between a raw data set and an estimated data set. In the raw data sets, missing observations and extreme values are unchanged; they are estimated in the estimated data sets. Median values estimate missing observations and extreme values. Both the raw and estimated data sets have three averaging periods.

Because daily data contains a high noise component that may mask a long–term cycle, and because many scientists are generally interested in long–term behavior of precipitation, the weekly, semi–monthly, and monthly behavior of each series is examined. Four averages per month comprise the weekly data. The first three averages include seven days each and the fourth contains the rest of the days in the month, so that a year contains forty-eight observations. Two averages for every month compose the semi–monthly data set; the first averages the beginning fifteen days of the month, and the second refers to the average of the remaining days. The monthly data set consists of a single average for each month. Median values replace missing or extreme observations because of the skewness of precipitation data. If the average

precipitation for the first week of May, 1990 is missing in the weekly data set, it is estimated with the median of the subsequent first weeks in May. Summary statistics of both the raw and estimated data sets for each averaging period are tabled for easy comparison.

Summary statistics, including the mean, median, standard deviation, and other statistics, give information about the location, skewness, and variability of the data. A graphical way to represent the same information is a line plot of the data. Line plots are high–density plots that, instead of graphing points on a two–dimensional plane, draw a vertical line from the horizontal axis to the ordinate. For a time series with many observations, line plots help identify any cyclic patterns, trends, or outliers more effectively than ordinary scatterplots. They lessen the masking behavior of noise, or random fluctuation, present in the data [12].

Almost any time series analysis requires second order stationarity (SOS). If a random process is SOS, it has a constant mean and variance over time and a covariance structure dependent only on the time lag between observations. One way to check for stationarity is to divide the observed time series into sub–series and examine the sub–series means and standard deviations for significant differences [5]. If the process that generated the observed time series is SOS, the means and standard deviations of the sub–series vary by chance. Each observed series was divided into fourths, then the resulting means were compared by analysis of variance and the homogeneity of the sub–series variances was tested using Cochran's test. It uses a ratio of sub–series standard deviations to determine significant differences between series variances. Both analysis of variance and Cochran's test assume independence of observations, an assumption that is violated if the overall series is anything but a randomly fluctuating sequence. However, these tests provide an indication of the stationarity of the series.

A least–squares regression line may also be used to determine whether an observed series has a linear trend. A scatterplot with the least–squares line and the mean of the series plotted as a horizontal reference line give a graphical representation for any observed trend. Although a t–test for the slope assumes independence and normality, a small p–value indicates any trend significance.

One goal of this study is to identify the cyclic components of each data set. One would suspect a yearly cycle, as it is widely assumed that weather varies with the seasons. However, many scientists believe weather patterns follow 10–year, 20–year, 100–year, and longer cycles, as well as shorter cycles with possibly two to four month periods. Neither of the available data sets contain enough information to identify a cycle longer than eighteen months, so this analysis concentrates on the detection of yearly and sub–yearly cycles.

An initial method of identifying a periodic component is through a scatterplot smooth. A commonly chosen smoother is the lowess function developed by William Cleveland [4]. It is an iterated, locally–weighted regression function that calculates a fitted, or smoothed, value of y at every time point by using a smoothing parameter chosen by the user. The smoothing parameter determines the fraction of surrounding data points that receive non–zero weights in the calculation of the fitted value [12]. Lowess follows an iterative procedure to avoid distortion caused by a small fraction of extreme values, creating a robust smooth. A trial–and–error process of under– and over–smoothing helps the analyst choose a final smoothing parameter to use for a given scatterplot. Under–smoothing results in a high–resolution but overly rough smooth, while over–smoothing results in a low–resolution smooth that does not respond to patterns in the data. The final smooth should be a compromise between the two extremes.

If a single periodic function generates the observed data, the peaks in the

scatterplot smooth appear nearly the same distances apart. For example, a monthly data set with an underlying yearly cycle will have smooth peaks nearly 12 observations apart. As the number of component frequencies increase, more irregularly–spaced peaks may result, making the lowess smooth more difficult to interpret.

Another plot used to investigate the periodic behavior of a time series is a smooth of the residuals from the series mean. A periodic series has periodic residuals. When graphed as a line plot, negative residuals appear as lines extending below zero and positive residuals as lines extending above zero. The smooth then serves as a guide to the eye when examining the patterns of positive and negative residuals for cyclic patterns. Once again, when a single underlying cycle is present in the data, the residual smooth will be nearly periodic. Then, as the number of component frequencies increases, the smooth becomes more difficult to interpret as component frequencies result in irregularly–spaced peaks.

Some authors suggest dividing an observed time series into sub–series by the length of the suspected period and then plotting each sub–series on the same set of axes. For example, since weather supposedly follows a twelve month cycle, a monthly series could be divided into several sub–series of length twelve. Each line on the graph would then represent a different year. If the series has an underlying twelve month component, each line will randomly fluctuate about the underlying cycle and a pattern emerges in the plot.

An alternative to the sub–series plot consists of a group of plots that show cyclic components without the noise component. Suppose a monthly series that spans seven years is divided into seven yearly series. Then each month has seven observations. Boxplots, median traces, and confidence intervals for each monthly median can be plotted side by side and their overall behavior examined for cycles. The side–by–side boxplots allow the interquartile ranges, minimums, and maximums to

be compared without other observations cluttering the plot. A median trace possibly helps determine a cycle without the cluttering of interquartile ranges, minimums, or maximums. This method plots the median from each boxplot and connects them with a line. The pointwise confidence intervals for the medians are line plots of the lower and upper 95% confidence limits for each median. Each interval has a confidence level of 95%, resulting in much lower simultaneous confidence levels. The width of the confidence intervals depends on the confidence level and on the number of observations available for the calculations. If a series does not divide equally, the resulting confidence interval widths differ because of variability within the observations' differing sample sizes.

Many times the component cycles in a time series cannot be identified by eye alone; either the noise element masks cyclic behavior or one cycle cannot be discerned from another. The correlogram and partial correlogram may identify the obvious and hidden correlations in the data, leading to previously undetected cyclic components. The correlogram and partial correlogram calculated by Splus are line plots [11]. The horizontal axis is lag and the vertical axis is the value of the correlogram or partial correlogram. The value at a given lag is represented by a line extending above or below zero to the value of the function at that lag. Splus also plots 95% error bars which correspond to $\pm\frac{2}{\sqrt{T}}$. Significant spikes extend above or below the error bars. Observations are negatively correlated at a given lag if the spike is significantly negative and vice versa for positive spikes. For example, if a monthly series has an underlying yearly cycle then observations twelve units apart are positively correlated. The correlogram and partial correlogram then show a significant positive spike at lag twelve. If observations six months apart are negatively correlated, the correlogram and partial correlogram show a significant negative peak at lag six. Any other lags at which the correlation between observations is significant correspond to significant

positive or negative spikes.

Univariate Time Series Analysis

Examination of the correlogram and partial correlogram plots leads to a list of candidate models for each of the six series. Each candidate model is considered by comparing the residual mean square and examining the significance of the estimated coefficients. If the model has more than one coefficient, the correlations among them are also examined. Running residual diagnostics gives an indication of model adequacy. Residual diagnostic plots include a line plot of the standardized residuals, a correlogram of the raw residuals, a plot of the p–values for the goodness of fit statistics, and a cumulative periodogram of the raw residuals. Another qualitative check involves comparing the theoretical spectrum of the fitted model to the smoothed periodogram. The model chosen has the simplest form, shows no correlation between coefficients, and passes all diagnostic checks.

Once a model is chosen, a residual is the difference between an observed value and the value of the model at a given time. Every observation has a corresponding residual, resulting in a residual time series the same length as the original series. An adequate model explains much of the observed variation between observations, so the residuals should resemble random fluctuation. If the residuals can be modeled as anything but white noise, the model is not adequate. A line plot of the standardized residuals allows for a visual check for patterns. The standardized residuals are used in the line plot to remove any non–constant variance exhibited in the raw residuals. Patterns in the standardized residuals usually signify model inadequacy.

The correlogram of the raw residuals is examined to check for autocorrelations in the residuals that may not be discernible in the line plot. An adequate model results in a correlogram with a positive spike at lag zero, and insignificant spikes at

all other lags.

The p–value for the portmanteau test statistic is calculated for several lags and plotted against lag as another diagnostic tool. A 5% significance line also appears on the graph for easy identification of any significant p–values. p–values falling below the 5% line signal possible inadequacies in the model. Sampling variability allows for a few p–values to fall below the significance line, but common sense and experience help the scientist determine any practical significance.

The cumulative periodogram of the residuals provides a diagnostic check that supplements the residual correlogram. If the residuals are white noise, the cumulative periodogram ordinates increase from 0 to 1 nearly linearly. 5% significance bands for the portmanteau test–statistic appear on the plot to guide the eye. A cumulative periodogram falling inside the significance bands provides evidence of a white noise sequence and an adequate model.

A final diagnostic uses the original data rather than the residuals. The theoretical spectrum can be derived for any ARMA model. A comparison between the theoretical spectrum of the fitted model and the smoothed periodogram provides a qualitative measure of model adequacy. Similar behavior of the two indicates that component frequencies in the data are included in the model.

If a candidate model fails any diagnostics, consideration of another model begins. Oftentimes the diagnostics of an inadequate model provide an indication of the next model to consider. The most desirable model is a simple and adequate one.

A major function of a time domain model is to forecast future values. Although a model describes the data to a certain extent, it doesn't go so far as to identify all periodic components. A particularly noisy data set can have an underlying periodic component that gets completely masked by noise and goes unidentified in an ARIMA model. Identifying hidden cycles is the prime function of frequency domain analysis.

Although the smoothed periodogram provides a diagnostic check for model adequacy, it also provides the strongest evidence of any underlying cycles.

Univariate Frequency Domain Analysis

Frequency domain analysis begins with an examination of the smoothed periodograms for each series. Large peaks in these plots help the analyst identify component frequencies of each time series. One expects that the three SNOTEL series will share the same component frequencies, and similarly for the three NOAA series.

One way to check the hypothesis of a yearly cycle is to calculate the period of a yearly cycle for each data set and add it as a reference line to the graph of the smoothed periodogram. The period of a yearly cycle has length 48, 24, and 12 for the weekly, semi–monthly, and monthly data sets, respectively. Sub–yearly cycles, such as a seasonal cycle, have shorter periods. The smoothed periodogram plots have reference lines corresponding to cycles with periods of 12, 6, 4, and 3 months plotted to guide the eye.

Because the periodogram is a smoothed estimate, the peaks may drift away from actual component frequencies as the smoothness increases. This makes identifying component frequencies somewhat subjective. The distance between a peak and its closest reference line must be evaluated by the analyst according to previous experience and results found in the exploratory and time domain analyses. In this situation, it is helpful that three different averaging periods for the same underlying series are being evaluated. The distances between the peaks and the reference lines may be evaluated in three different instances in order to make a sound conclusion about which frequencies are important.

Bivariate Frequency Domain Analysis

Once the SNOTEL and NOAA series have been evaluated individually, it is important to consider them together to gain an overall understanding of their relationship. The goal here is to understand how similarly the two series behave: the relative importance of component frequencies in each and how the two cycle together. A bivariate time domain model would not necessarily answer these questions, so only the frequency domain is considered in this paper.

Bivariate frequency domain analysis begins by considering only the observations from each series that involve the intersection of the time domains for each location. The two reduced series are called the common time series. A plot of the common series helps determine qualitatively whether they behave similarly across time. One way to plot the two series is to add a constant to every observation of one series so that when the two are graphed on the same axes they don't overlap. This method reduces the confusion introduced when one highly erratic line is plotted over another.

Another way to get a feeling for the relationship across time is to plot the datasets as paired observations. The two series may be considered a set of paired observations $(x_t, y_t), \quad t = 1, 2, \ldots, T^*$ where T^* represents the intersection of the time domains of each series. A reference line $x = y$ can be added to the plot along with a lowess smooth to guide the eye when looking for an overall pattern of similar behavior.

Once the general behavior of the common series in the time domain is examined, the smoothed marginal periodograms of the common series may be examined for each averaging period. The periodograms are overlayed on the same axes for easy comparison. If the periodograms from the two show no common component frequencies, the analysis is essentially complete with the conclusion that the two do not cycle

together, so share no common behavior.

When the two peak in the same bands of frequencies, the next step involves determining the relative importance of the shared component frequencies and estimating any phase shift between corresponding components. The coherence spectrum provides a measure of the relative importance of a component frequency as an expression of the squared correlation between amplitudes. The phase shift measures the difference in phases between series' components.

Because coherence actually measures the squared correlation between the amplitudes at a specified frequency and is bounded by the interval $[0, 1]$, values of zero indicated no correlation, while values of one indicate perfect correlation. Series may actually cycle together with little correlation between amplitudes. For this reason, one is especially interested in frequencies where the marginal periodograms peak together and show high coherence.

Once frequencies of interest are identified, the phase shift between can be examined for each averaging period. The ordinates in the estimated phase spectrum are standardized so they fall into the interval $[0, 2\pi)$. This allows for easy transformations of the phase shift into the time domain. For example, phase estimates near 0 or 2π indicate no phase shift, while values near π indicate a shift of half a period. A constant phase shift at all frequencies indicates one series leads the other in terms of precipitation events.

CHAPTER 5

SNOTEL DATA ANALYSIS

Exploratory Data Analysis

The SNOTEL data set begins October 1, 1984 and ends February 29, 1992. It contains a single large block of missing observations: the year–long period from October 1, 1989 through September 30, 1990. Other occasional missing data points occur but no large blocks. There are two observations greater than 1.5 inches, December 15, 1987 with 2.2 inches and May 8, 1988 with 2.0 inches. These observations contribute to two of the outliers seen in the weekly plots in Figure 1 but they get absorbed into the means in the semi–monthly and monthly data sets. Precipitation is measured in tenths of inches.

Table 1: Summary statistics for the SNOTEL data sets

	raw series			estimated series		
	weekly	semi–monthly	monthly	weekly	semi–monthly	monthly
mean	0.08	0.08	0.08	0.08	0.08	0.08
st.dev.	0.08	0.07	0.05	0.08	0.06	0.05
min	0.0	0.0	0.0	0.0	0.0	0.0
median	0.07	0.08	0.08	0.06	0.08	0.08
max	0.59	0.44	0.26	0.59	0.44	0.26
n	305	153	77	353	177	89

Table shows the summary statistics for the various SNOTEL data sets. There are six data sets: weekly, semi–monthly, and monthly in which missing dates are ignored, and weekly, semi–monthly, and monthly where values for the missing dates have been estimated as described in the **METHODS** section. The six data sets are divided into two groups: those with the missing dates ignored, referred to as the raw

data sets, and those in which missing observations are estimated, called the estimated data sets.

Figure 1 shows line plots for the six SNOTEL data sets. The first column of plots shows the raw data and illustrates the locations of the missing observations. The second column shows plots of the corresponding estimated data sets. The weekly plots show four relatively extreme observations. They occur during the third week of February, 1986; the third week of December, 1987; the second week of May, 1988 and the first week March, 1991. Their respective readings are 0.59, 0.47, 0.40, and 0.39 inches. As the averaging time increases, the extreme observations' influence diminishes. In the monthly plots, the extreme observations seem to meld in to the cyclic patterns noticeable in those plots.

The line plots also illustrate two particularly dry periods: the last couple of months of 1987, and the last six months of 1988. Most people, especially those in Montana and Wyoming, recall the 1988 dry period, as it corresponds with the Yellowstone Park fires that were burning at that time.

The cyclic behavior becomes easier to identify as the time averages increase. In the weekly plots, the seasonal pattern gets masked by the noisy nature of the data. This noise factor diminishes as the averages include more observations. The monthly plots show lower precipitation in the middle portions of each year and higher precipitation readings toward years' beginnings and endings.

Figure 2 shows plots of each data set divided in fourths so that the means and standard deviations of the resulting sub–series can be tested for differences. If any of the data sets show significant differences among the means or standard deviations, the data may not be stationary. The means were compared by an analysis of variance(ANOVA) and the standard deviations by Cochran's test. The ANOVA p–values and the Cochran's test–statistics and critical values for the $\alpha = .01$ level are listed