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 — William F. Quimby

Date — William F. Quimby
Chairperson, Graduate Committee

Approved for the Major Department

4/5/95 — John Lund

Date — John Lund
Head, Statistics

Approved for the College of Graduate Studies

4/23/95 — R L Brown

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 Cah

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) random 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 $\hat{r}_j$ estimates the j^{th} seasonal coefficient.

If the residuals are independent at the first two regular lags, i.e. $\mathrm{E}\left[e_t, e_{t+1}\right] = \mathrm{E}\left[e_t, e_{t+2}\right] = 0$, 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 $\mathrm{E}\left[e_t, e_{t+S}\right] = \mathrm{E}\left[e_t, e_{t+2S}\right] = 0$, 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 $C_j = \sum_{k=1}^{j} I(\alpha_j) : \quad j = 1, 2, \ldots, m$. 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_{xy}^2, which can be written

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

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

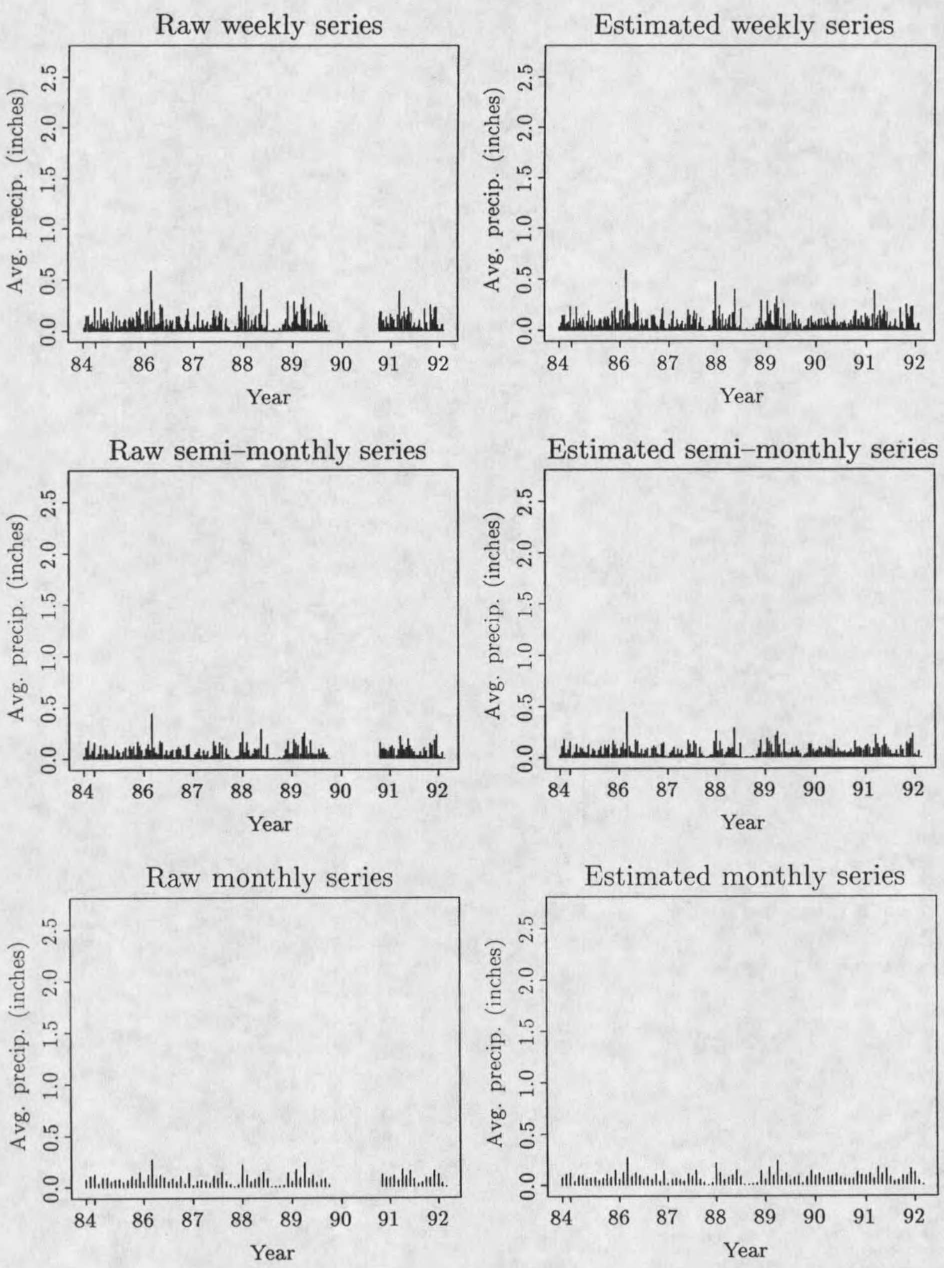

Figure 1: Line plots of the SNOTEL raw and estimated data sets for weekly, semi–monthly, and monthly averaging periods.

in Table 2. The table contains results from three different sets of tests. The within series–raw and within series–est portions give results for tests on the individual data sets while the between series portion gives results for the comparison of means and standard deviations of the raw and estimated data sets combined. Any significant results for the third series of tests would indicate the raw and estimated data sets differed in a statistically significant way.

	Within series–raw			Within series–est		
Averaging period	ANOVA p–value	Test–statistic	Critical value	ANOVA p–value	Test–statistic	Critical value
weekly	0.348	1.65	2.3	0.393	1.306	2.3
semi–monthly	0.445	1.361	3.3	0.520	1.96	3.3
monthly	0.636	1.44	4.3	0.742	2.25	4.3

	Between series		
Averaging period	ANOVA p–value	Test–statistic	Critical value
weekly	0.680	1.65	1.96
semi–monthly	0.843	1.96	2.63
monthly	0.982	2.25	3.32

Table 2: ANOVA p–values for differences between sub–series means and Cochran's test–statistics and critical values for homogeneity of variance between sub–series illustrated in Figure 2.

Because there are no significant differences between means or standard deviations within series, second order stationarity may be assumed for both the raw and estimated series. Estimating the missing observations does not affect stationarity because there are no significant differences between means or standard deviations between sub–series. This provides one indication that the estimated series may be substituted for the raw series.

Lowess smooths of the SNOTEL data sets are shown in Figure 3. The first column shows smooths of the raw data with the missing year removed. If the SNOTEL data is generated by an underlying yearly cycle, the removal of an entire year

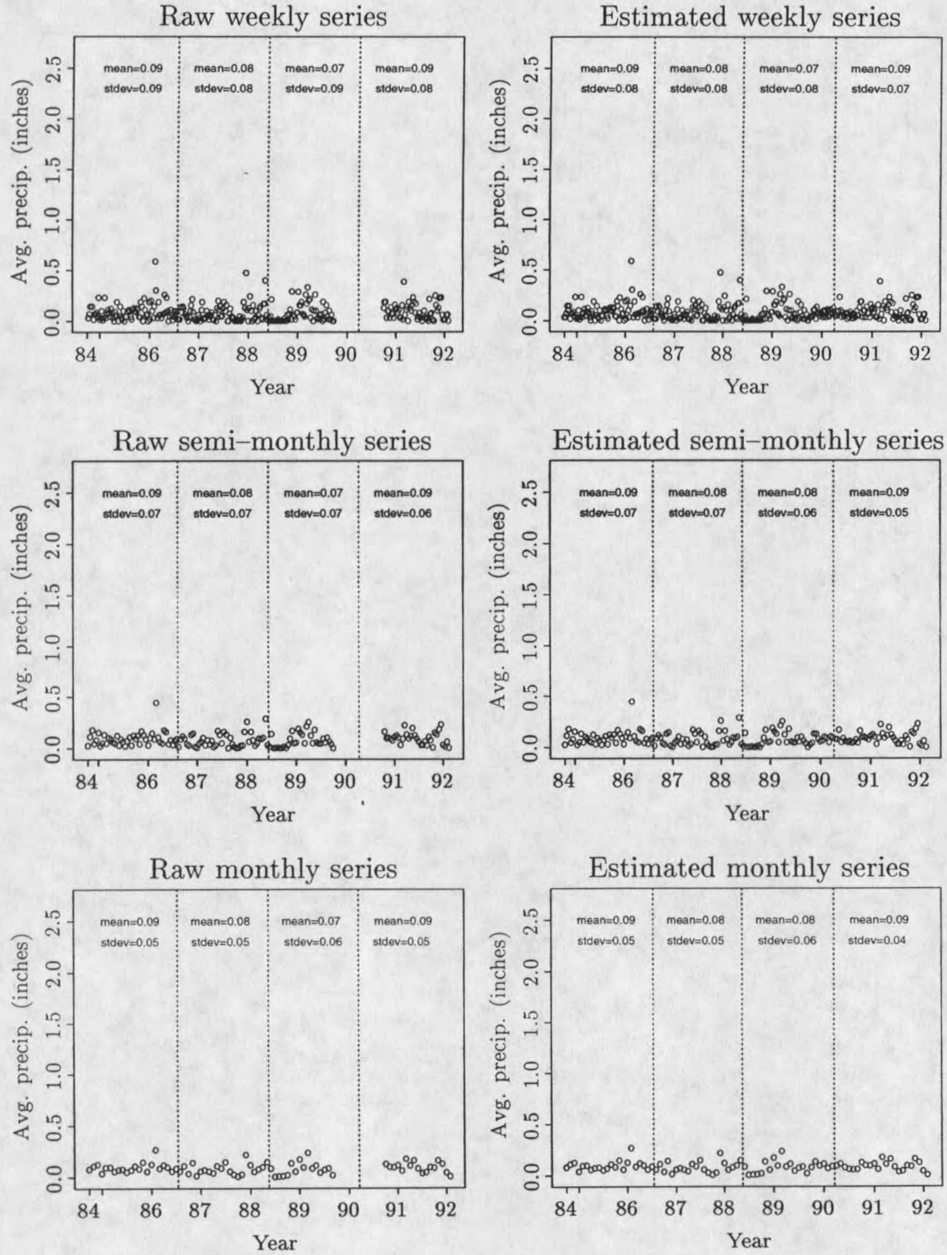

Figure 2: SNOTEL raw and estimated weekly, semi–monthly, and monthly datasets are divided into fourths. Means and standard deviations for each sub series are included.

should not affect any patterns in the data. The second column shows smooths of the estimated SNOTEL data sets.

Because each smooth is calculated with a 0.10 smoothing parameter, corresponding raw and estimated data sets have identical smooths until the beginning of the missing data. They also appear similar after the period of missing data. The only difference between the raw and estimated smooths is an extra peak in the estimated smooth for the time with estimated observations. The oscillation of the smooths indicates a cyclic pattern with approximately one peak per year, but the peak seems to shift in location from year to year. The missing year follows that pattern with a single peak, providing evidence for a yearly cycle of precipitation at the SNOTEL site. The shift in the peak from year to year indicates that another cycle, perhaps with a 6–month period, also contributes to the fluctuation.

Figure 4 shows line plots of the residuals from the overall mean. Again, the raw data residuals are shown in the first column and the residuals of the estimated data in the second. The residual smooths for both data sets are calculated with the same smoothing parameter but the raw data sets are shorter. Fewer data points in the smoothing window of the raw data causes a rougher smooth. Therefore, the raw smooths appear slightly rougher than their estimated counterparts. However, the same cyclic patterns emerge in each plot. As in Figure 3, the smooths indicated a yearly cycle, peaking approximately once a year.

Figure 5 shows the estimated SNOTEL data plotted in a yearly fashion. The horizontal axis corresponds to a twelve month period and each year of the series appears on the same axes. Only the estimated data is shown because the absence of a single year is indistinguishable in these plots. The monthly plot provides the strongest evidence of a yearly cycle. It shows a pattern of higher precipitation through the winter and spring, low precipitation in the summer, and a rise again in the late fall.

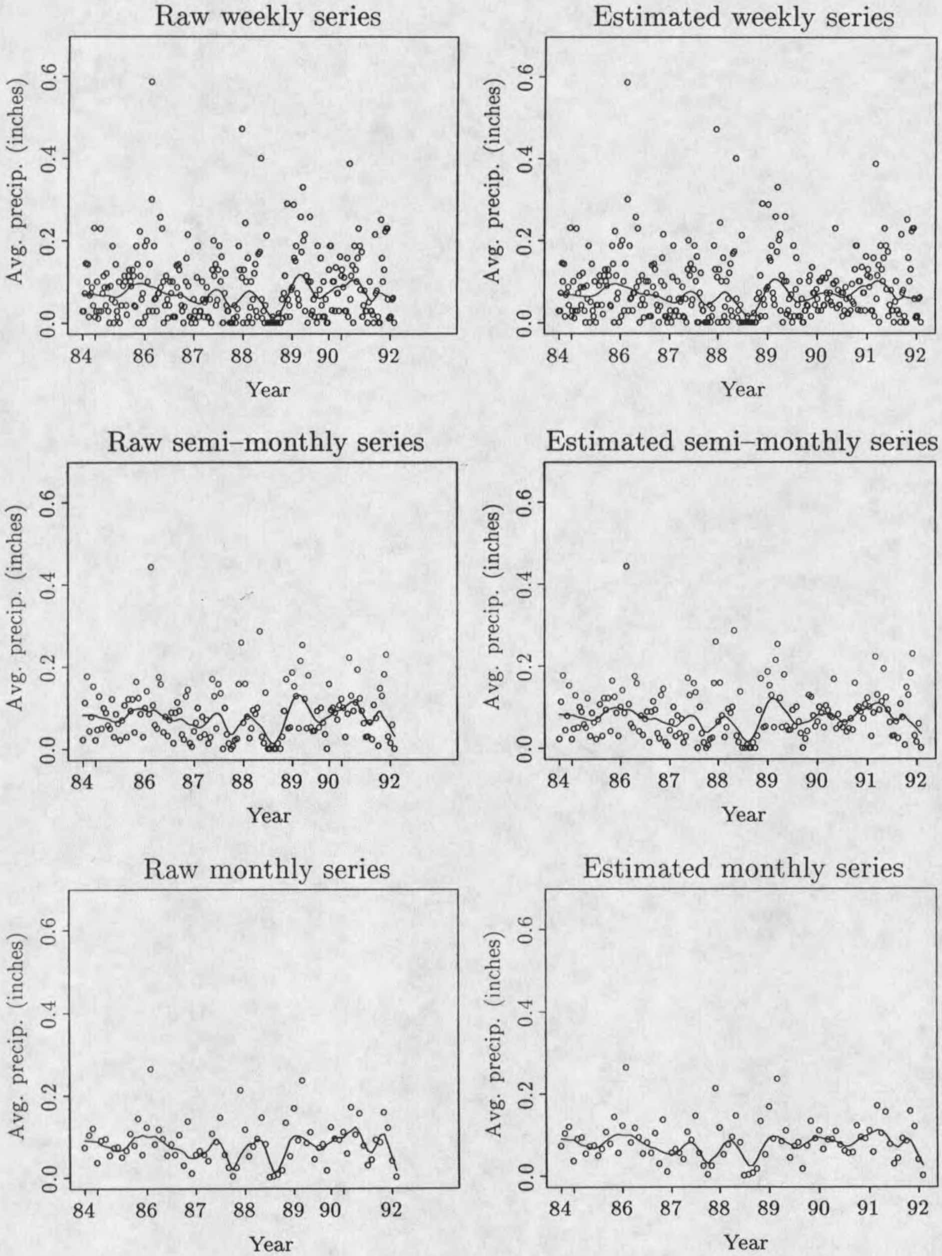

Figure 3: Lowess smooths of raw and estimated SNOTEL series with a smoothing parameter of $f = .1$.

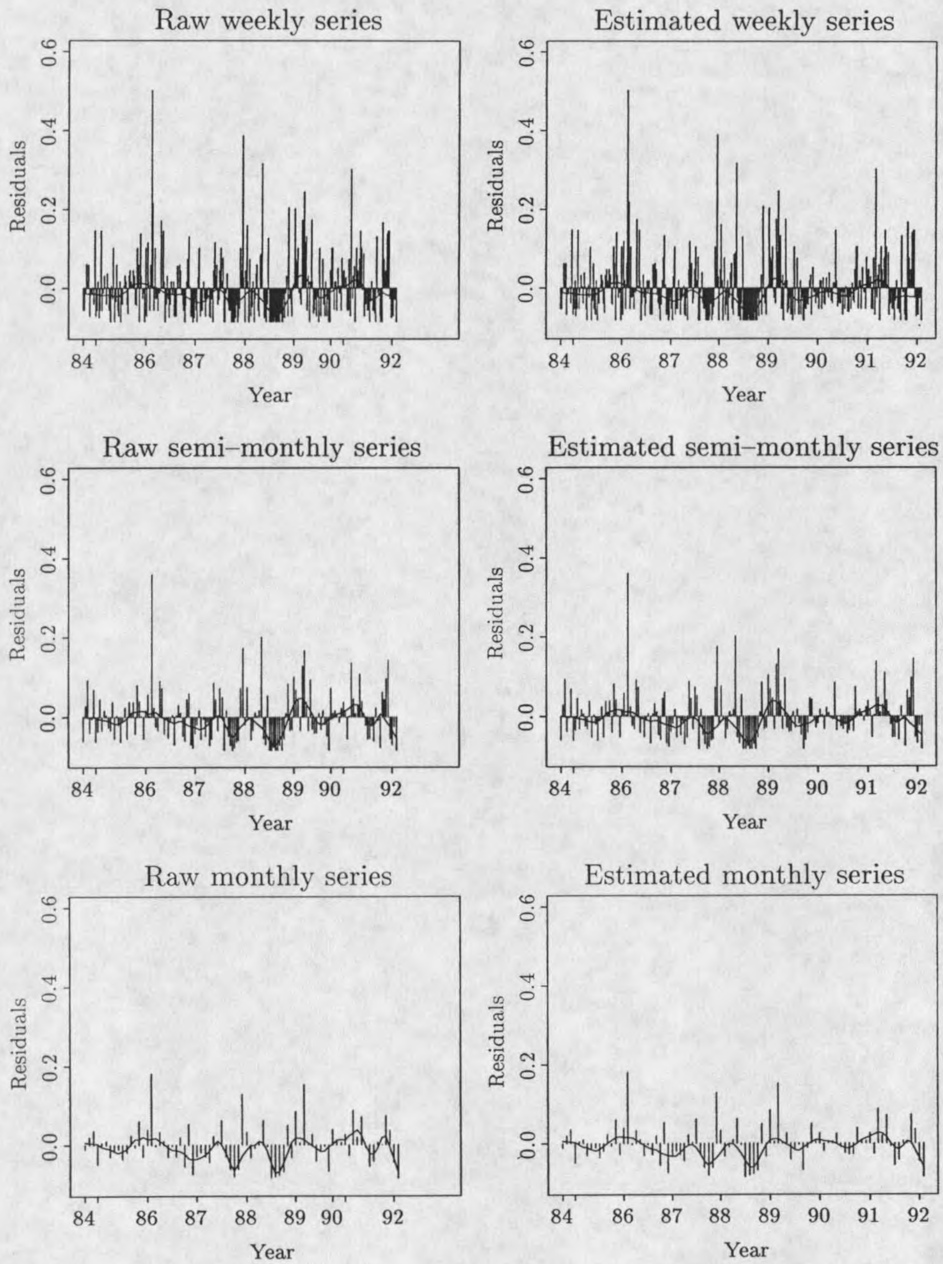

Figure 4: SNOTEL residuals from the mean for weekly, semi–monthly, and monthly data sets. The lowess smoothing parameter is $f = .1$.

However, it's difficult to determine if the pattern is strong. Two lines showing years of particularly high winter precipitation and two lines showing years with unusually low summer precipitation confuse the issue. The noisy nature of the data prevents this plot from providing overwhelming evidence of a yearly pattern.

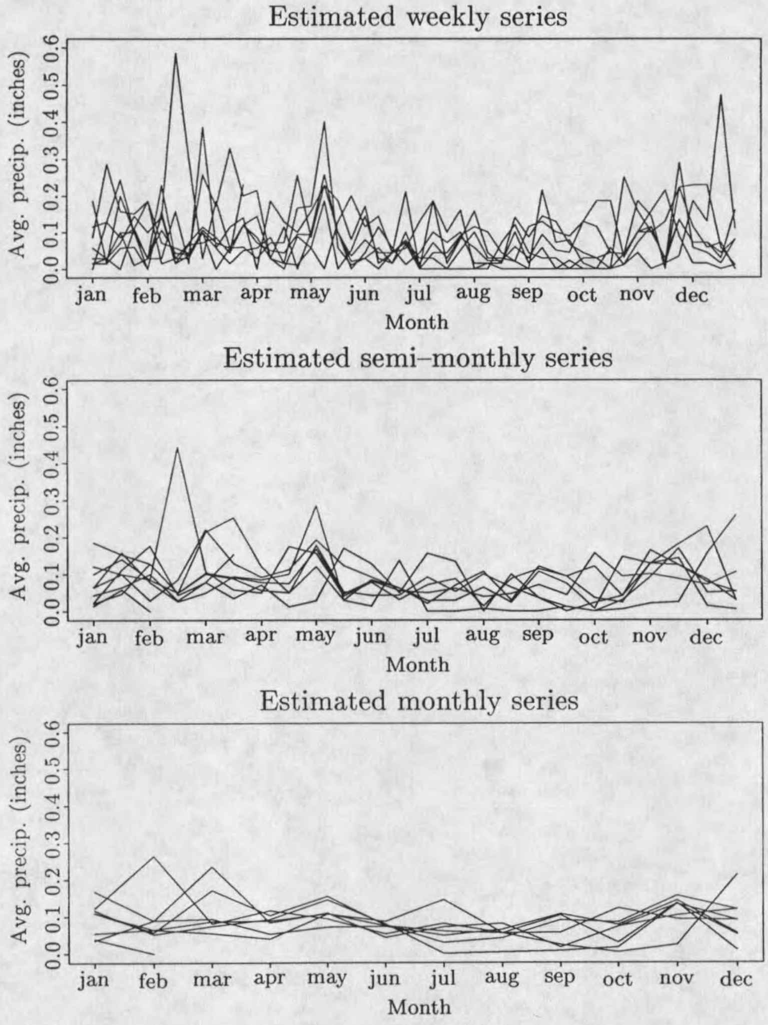

Figure 5: SNOTEL weekly, biweekly, and monthly estimated data sets. Each line represents a year of average precipitation values.

Alternatives to Figure 5 include a series of side–by–side boxplots, median traces, and pointwise confidence intervals for the medians, all of which are shown in Figures 6, 7, and 8. Boxplots illustrate the spread and skewness of the data while the median traces summarize the period–to–period behavior. The 95% confidence intervals for the medians most effectively illustrate the cyclic behavior, if any, present in the data. The boxplots, median traces, and confidence intervals for the raw data appear in the first column, their estimated counterparts in the second. Figure 6 shows weekly plots; Figure 7, semi–monthly plots; Figure 8, monthly plots. Similarities and differences between the raw and estimated data are discussed before the general behavior of the series is examined across time intervals.

Figure 6 reveals subtle differences between the raw and estimated data sets. The median traces of each look nearly identical, but there are a few differences in the boxplots and confidence intervals. Just as the means change little between the raw and estimated data, as illustrated in Figure 2, the same holds true with the medians. The lengths of the boxplots and confidence intervals change because of the increase in the number of observations from the raw to estimated data sets. There is one more observation for every period in the estimated data, but none of those additional observations are extreme. Therefore the boxplots and confidence intervals for the estimated data are generally shorter than the raw data's. This is similar to the decrease in the standard deviations noted in Figure 2. Other than the slight decrease in the lengths of the boxplots and confidence intervals, the plots for the raw and estimated data remain very similar.

These figures are also useful in describing the general behavior of the series across time intervals. First, the boxplots illustrate the skewed nature of the data: the medians are generally closer to the bottoms of the boxplots and the top whiskers are generally longer than the bottom ones. The largest outliers occur in the spring and winter months, and the outliers that occur in the summer months fall within the interquartile ranges(IQR's) of observations at other time periods, indicating even the

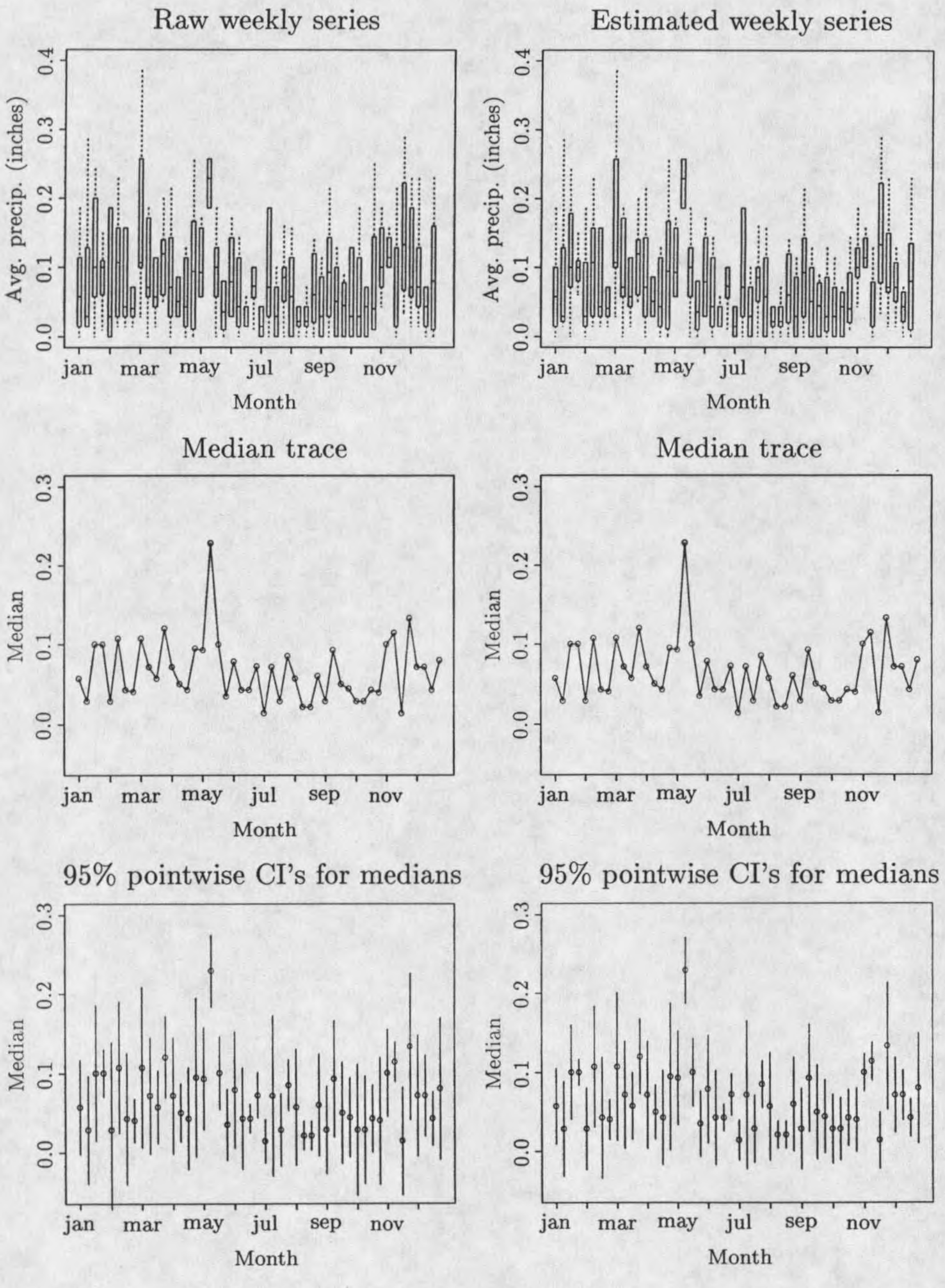

Figure 6: Boxplot, median trace, and 95% confidence intervals for the medians for SNOTEL weekly raw and estimated series. Outliers are omitted.

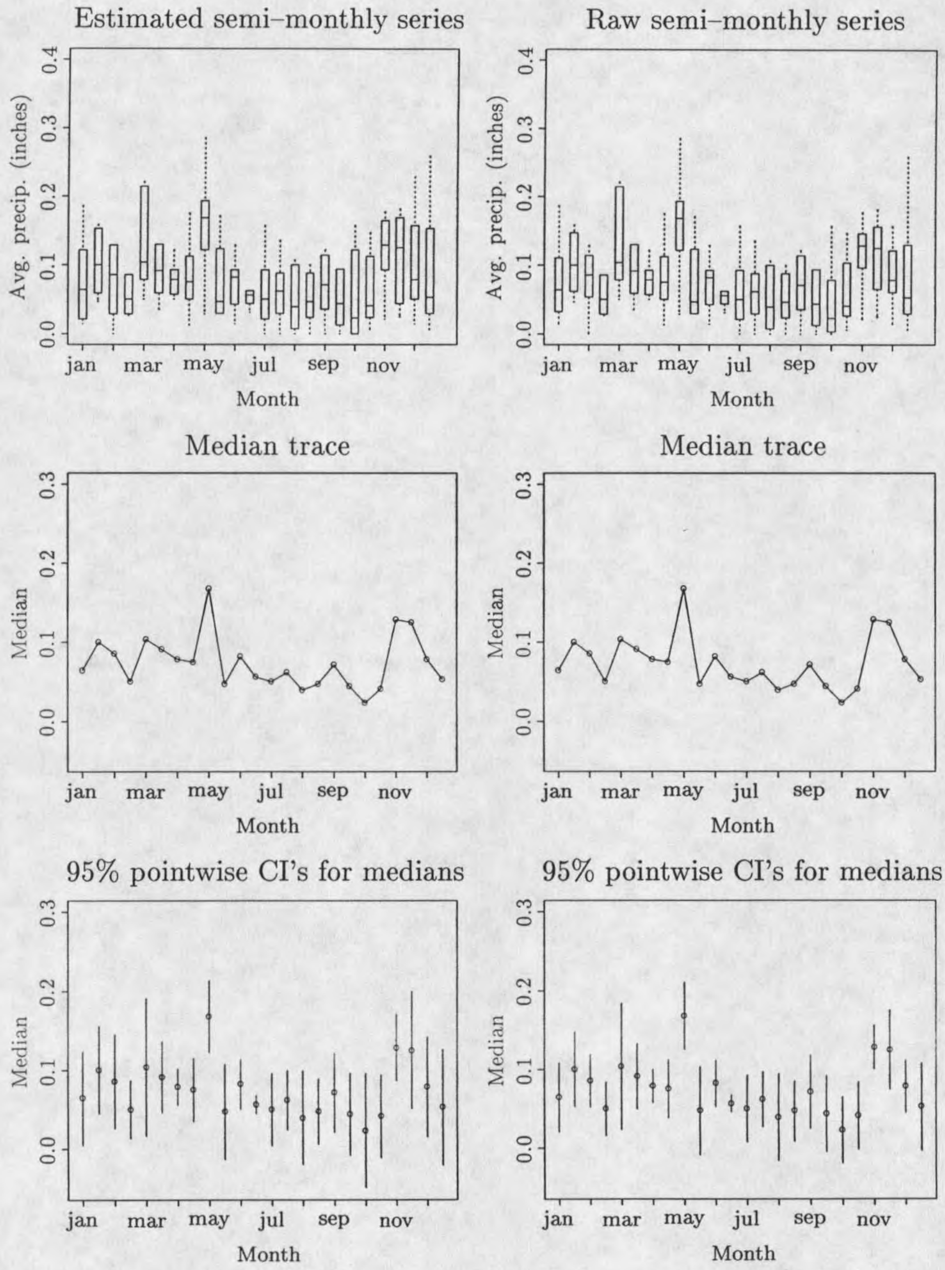

Figure 7: Boxplot, median trace, and 95% confidence intervals for the medians for SNOTEL semi–monthly raw and estimated series. Outliers are omitted.

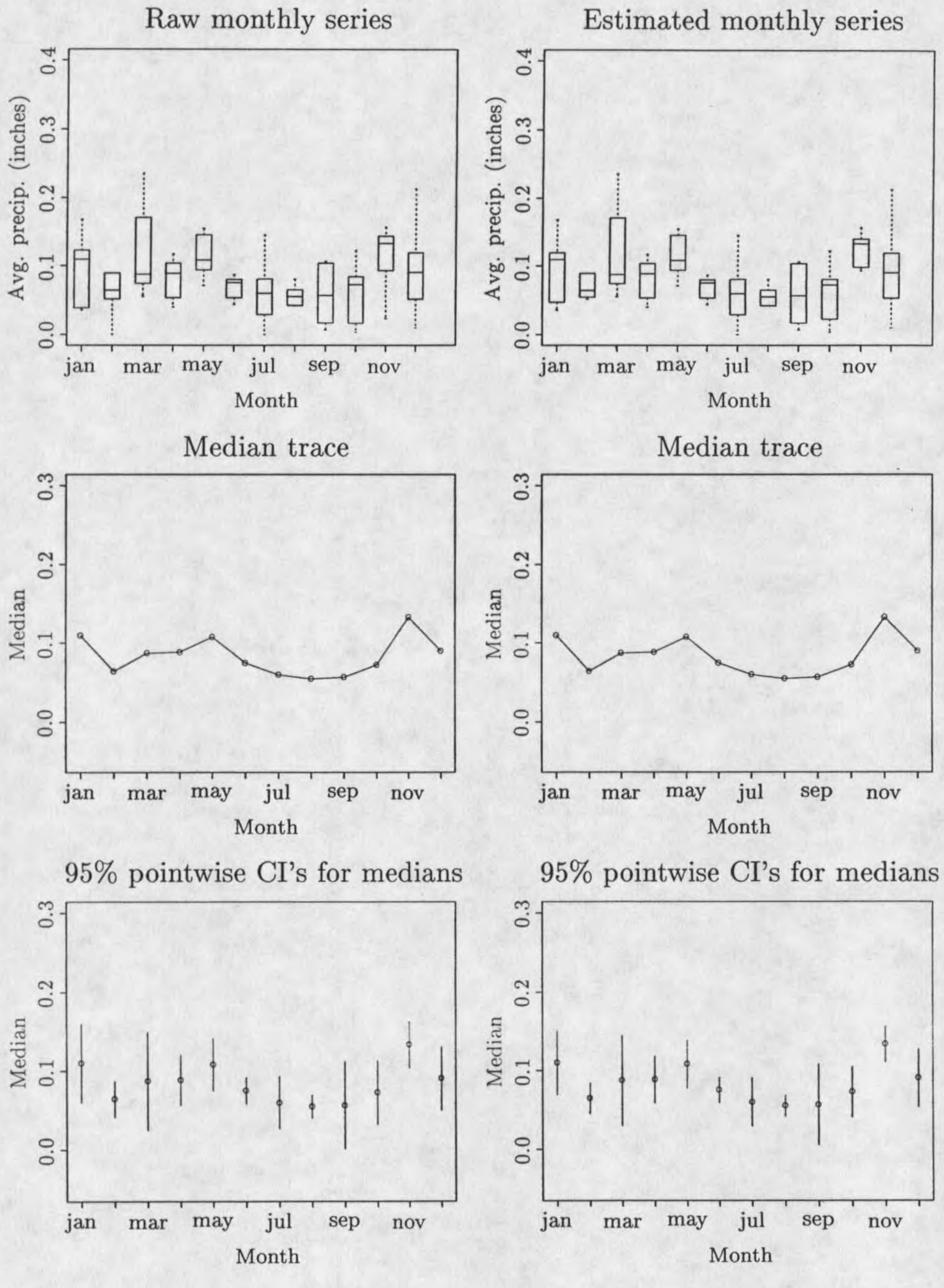

Figure 8: Boxplot, median trace, and 95% confidence intervals for the medians for SNOTEL monthly raw and estimated series.

extreme amounts of precipitation in the summer are within normal yearly ranges.

Second, the median plots show a yearly summary of precipitation by period. A yearly pattern emerges in the monthly median trace but is not as pronounced in the other traces; the weak pattern shown in the monthly trace is masked by noise in the other traces. The pattern illustrated in the monthly median trace is similar to the weak pattern shown in the monthly plot in Figure 5: somewhat low precipitation throughout the summer and higher amounts in the winter, spring, and late fall.

Third, the 95% pointwise confidence intervals for the medians reiterate the cyclic behavior found in the monthly median traces. Once again, the cycle is difficult to see in the weekly intervals because of the variability of the medians and the corresponding variability in the widths of the confidence intervals. The pattern of low summer precipitation emerges in the semi–monthly intervals, and is most distinguishable in the monthly intervals. As in the median traces, the confidence interval for May is higher than the others in both the weekly and semi–monthly plots.

Time Domain Analysis

In many instances, a series is too noisy to pick out cycles and correlation structures by eye. Plots of the sample autocorrelation and partial autocorrelation functions provide a complement to plots of the data for determining the cyclic nature of a time series. Figures 9 and 10 show the correlogram and partial correlogram of the raw SNOTEL data in the first column and the corresponding estimated data in the second column.

A comparison of the peaks in each acf shows that estimating the entire missing year has no effect on the dependencies within the original dataset. The peaks change slightly, but because none of the peaks are significant, the slight changes are irrelevant.

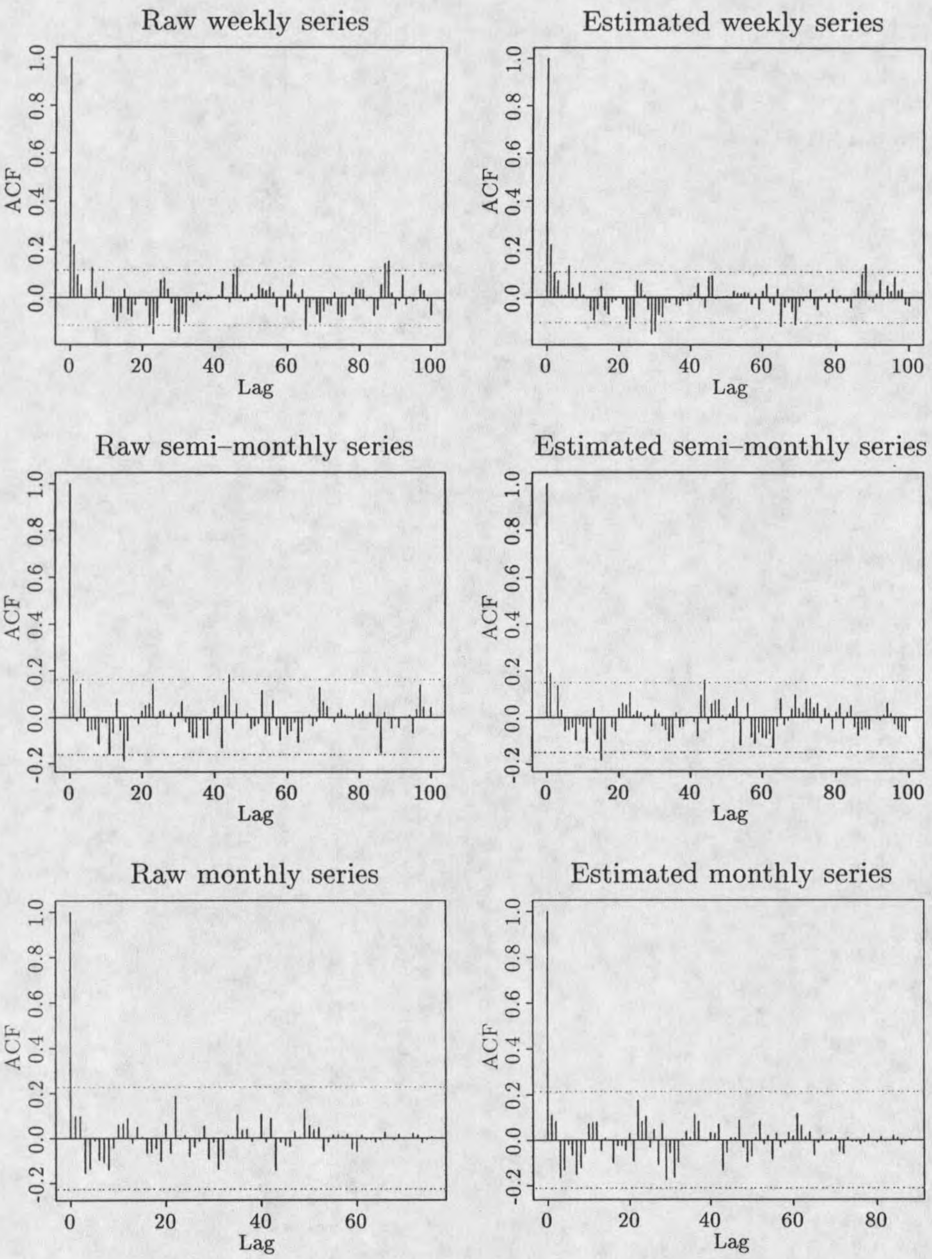

Figure 9: SNOTEL sample autocorrelation functions for weekly, semi–monthly, and monthly series.

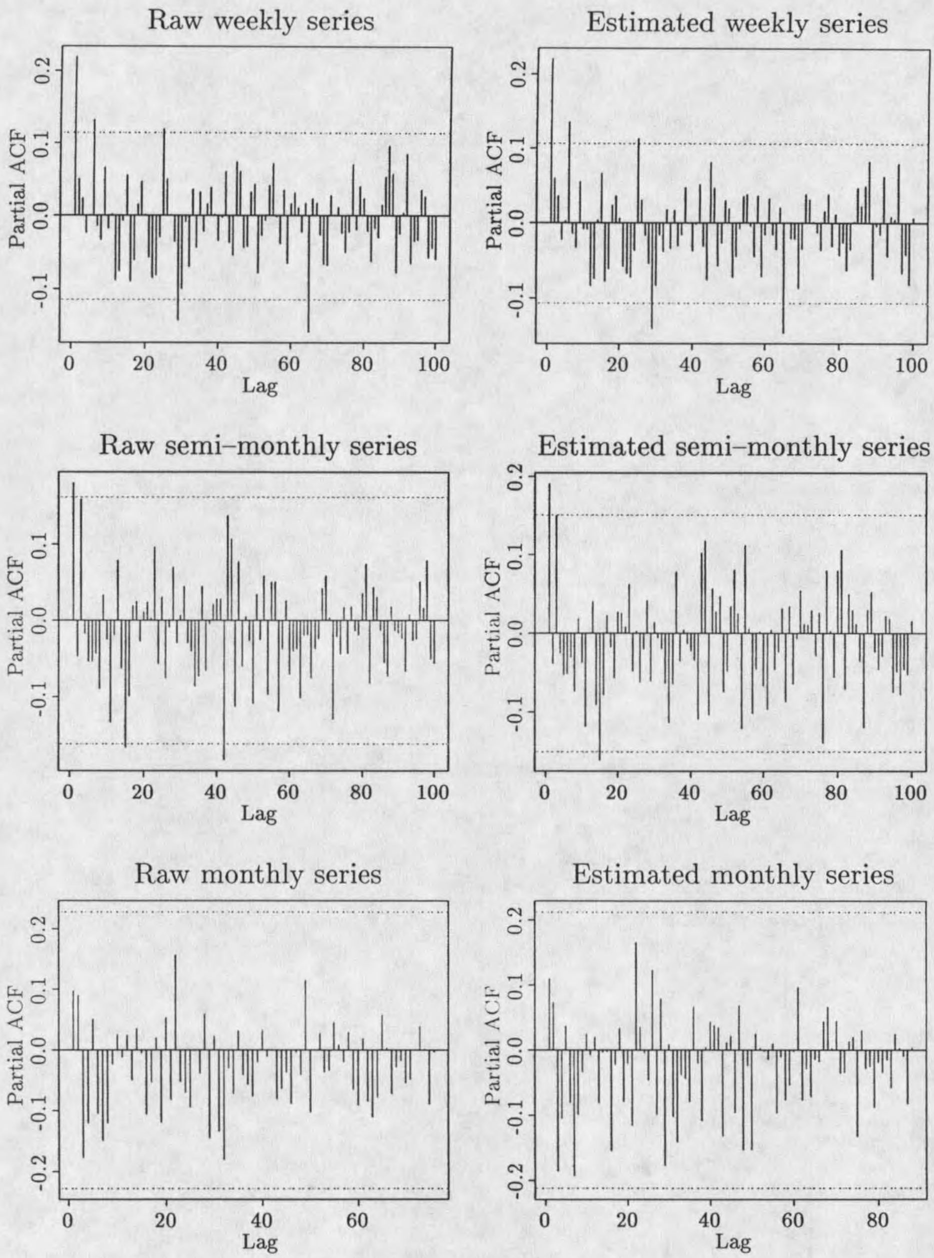

Figure 10: SNOTEL sample partial autocorrelation functions for weekly, semi–monthly, and monthly series.

According to the correlograms of the SNOTEL data, the correlation structure remains unchanged when the missing year is estimated.

The correlograms also illustrate little or no correlation between observations within the SNOTEL data set. Each weekly acf has a slightly significant peak at lag 1, indicating possible serial positive correlations among the weeks. However, the severe drop between lag 0 and lag1 autocorrelations indicates marginal significance, if any.

An interesting note here is the lack of significant spikes at yearly lags. The fact that lags corresponding to observations a year apart do not have significant acf values indicates the lack of a strong yearly cycle. If there were a strong yearly cycle, one would expect a significant spike at or near lag 48 in the weekly acf, lag 24 in the semi–monthly acf, and lag 12 in the monthly correlogram. The weak cycle discussed earlier is not strong enough for the acf to pick up; however, the pattern may still exist.

The partial correlograms are shown in Figure 10. As with the correlogram, the raw and estimated partial correlograms show only minor differences. The most notable feature between plots is that as the averaging period increases, the behavior of the partial correlogram quickly diminishes to that of a white noise sequence. Both weekly partial correlograms show positive spikes at lag 1 with questionable spikes at lags 6, 28, and 60. An $AR(1)$ may be a sufficient model for the weekly series. The semi–monthly partial correlograms show a barely significant spike at lag 1, indicating and $AR(1)$ at the most, and possibly a white noise model. The monthly partial correlograms show no spikes at any lag, clearly indicating white noise behavior. Recall that the pacf shows correlations between observations with the effects of all other lags removed. The monthly partial correlograms show that any correlation between observations has been buried in the average.

The similarities between the raw and estimated correlograms and partial correlograms indicates that the series can be modeled identically. Estimation of the missing year results in the same model as deletion. Figures 9 and 10 offer no evidence of a cyclic component in precipitation at the SNOTEL site. The only strong candidate for any averaged period is an $AR(1)$. This makes selecting a model rather straight forward.

The $AR(1)$ fit for the weekly series returns a coefficient of $\hat{a}_1 = .2209$ and $RMS = 18.762$. A t–test for $H_o : a_1 = 0$ on 352 degrees of freedom gives a p–value of .0000, indicating $\hat{a}_1$ is significant. Diagnostic plots appear in Figure 11.

The first plot shows a line plot of the standardized residuals. The skewed nature of the data is once again illustrated but no cyclic pattern appears. The correlogram of the raw residuals shows no significant spikes and none of the p–values for the goodness–of–fit statistic are significant. The cumulative periodogram of the residuals falls entirely inside the 95% critical bands indicating a white noise residual series. The theoretical spectrum of an $AR(1)$ process with coefficient .220885 appears as the dashed line while the smoothed periodogram of the data is the solid line. The similarity between the two reinforces the adequacy of the model.

Neither the $AR(2)$ nor the $MA(1)$ models reduce the RMS and the $AR(2)$ coefficient is not significant. Thus, the chosen model for the weekly SNOTEL series takes the form $Y_t = .2209Y_{t-1} + \varepsilon_t$.

The $AR(1)$ fit for the semi–monthly data returns a coefficient of $\hat{a}_1 = .1920$ and $RMS = 13.2665$. The p–value for a t–test with 175 degrees of freedom is .0102. Figure 12 shows the various diagnostic plots. The standardized residuals show some cyclic behavior, although the correlogram, goodness of fit plot, and cumulative periodogram support the model. The similarity between the fitted $ARMA$ spectrum and the raw periodogram of the data also supports the $AR(1)$ model.

Because the correlogram of the semi–monthly series contains a significant spike

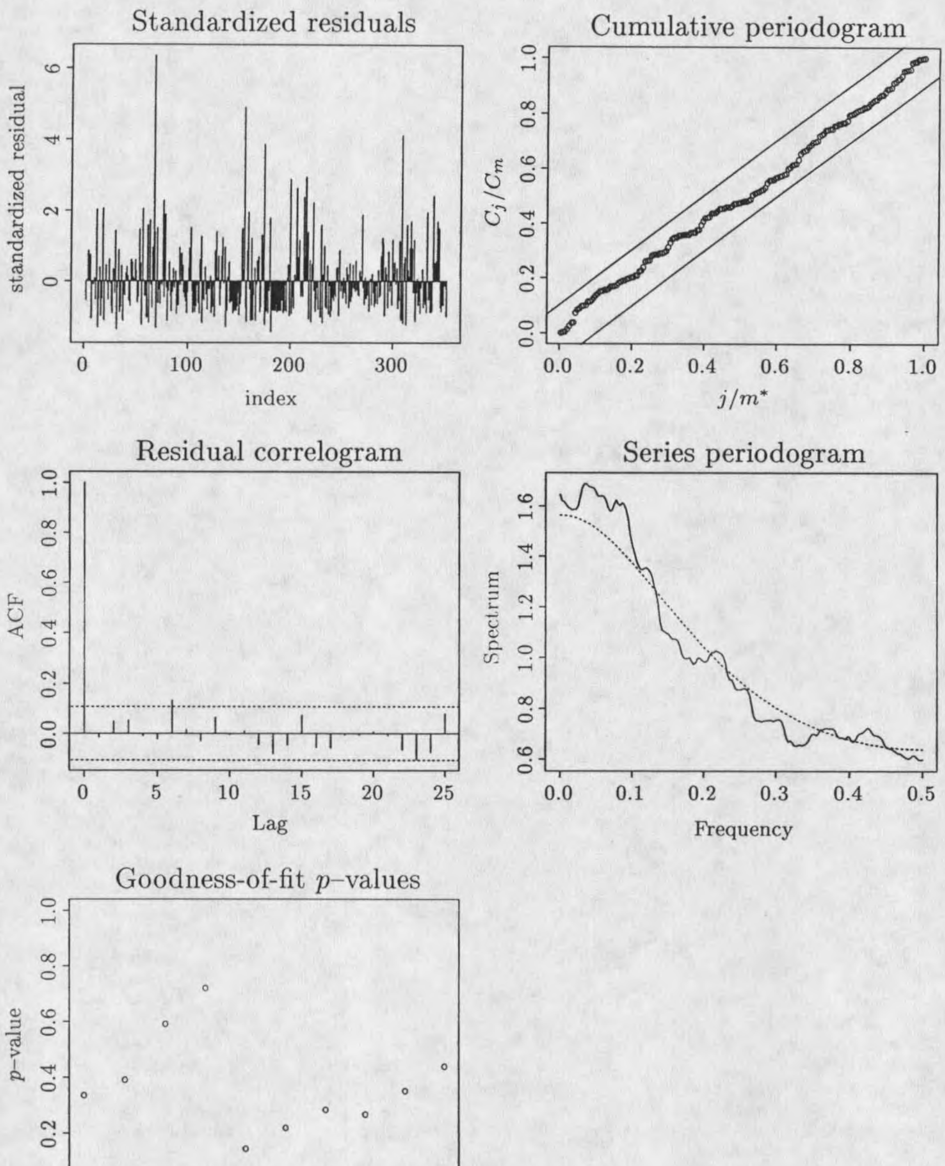

Figure 11: Diagnostic plots for the weekly SNOTEL $AR(1)$ fit.

at lag 15, an $AR(1)(1)15$ is also fit. This corresponds to a $7\frac{1}{2}$ month correlation between observations. The coefficients for this model are $\hat{a}_1 = .1948$ and $\hat{a}_15 = -.1847$. The correlation between coefficient is approximately -1.05×10^{-10}–small enough to consider the coefficients uncorrelated. The RMS for this model is 12.6885, a .578 decrease from the $AR(1)$ model. A t–test with 159 degrees of freedom for $H_o : a_{15} = 0$ gives a p–value of .0182.

The minimal decrease in the residual mean square and the significant p–value indicate the $AR(1)(1)15$ model is overfit. The $AR(2)$ and $MA(1)$ models give RMS of 13.2288 and 13.304. The plots and overfits verify the adequacy of the $AR(1)$ model. It looks like $Y_t = .1920Y_{t-1} + \varepsilon_t$.

The correlogram and partial correlogram for the monthly series show no significant spikes, therefore a white noise model may be appropriate. An additional check of this is to subtract the mean from each observation in the series and treat the resulting set as a residual series. A cumulative periodogram of these residuals provides a diagnostic for the white noise model. Figure 13 shows the cumulative periodogram of the monthly mean–removed SNOTEL series. The ordinates all fall within the 95% tolerance limits, supporting a white noise model.

Trial overfits for the white noise model include serial($AR(1)$) and purely seasonal models. The coefficient for an $AR(1)$ is $\hat{a}_1 = .1123$. The fit returns a residual mean square of 9.3808. A t–test for a_1 on 86 degrees of freedom gives a p–value of .2894, indicating an $AR(1)$ is indeed an overfit. The test–statistic in equation (2.5) with $v = 35$ gives a p–value of .0000.

The pure seasonal model is also fit in order to test the significance of the seasonal coefficient without a serial term. The fit returns a coefficient of $\hat{a}_{12} = .854$ and a residual mean square of 8.775. t–test for a_{12} on 76 degrees of freedom returns a p–value of .3954. Because the correlogram and partial correlogram show no significant

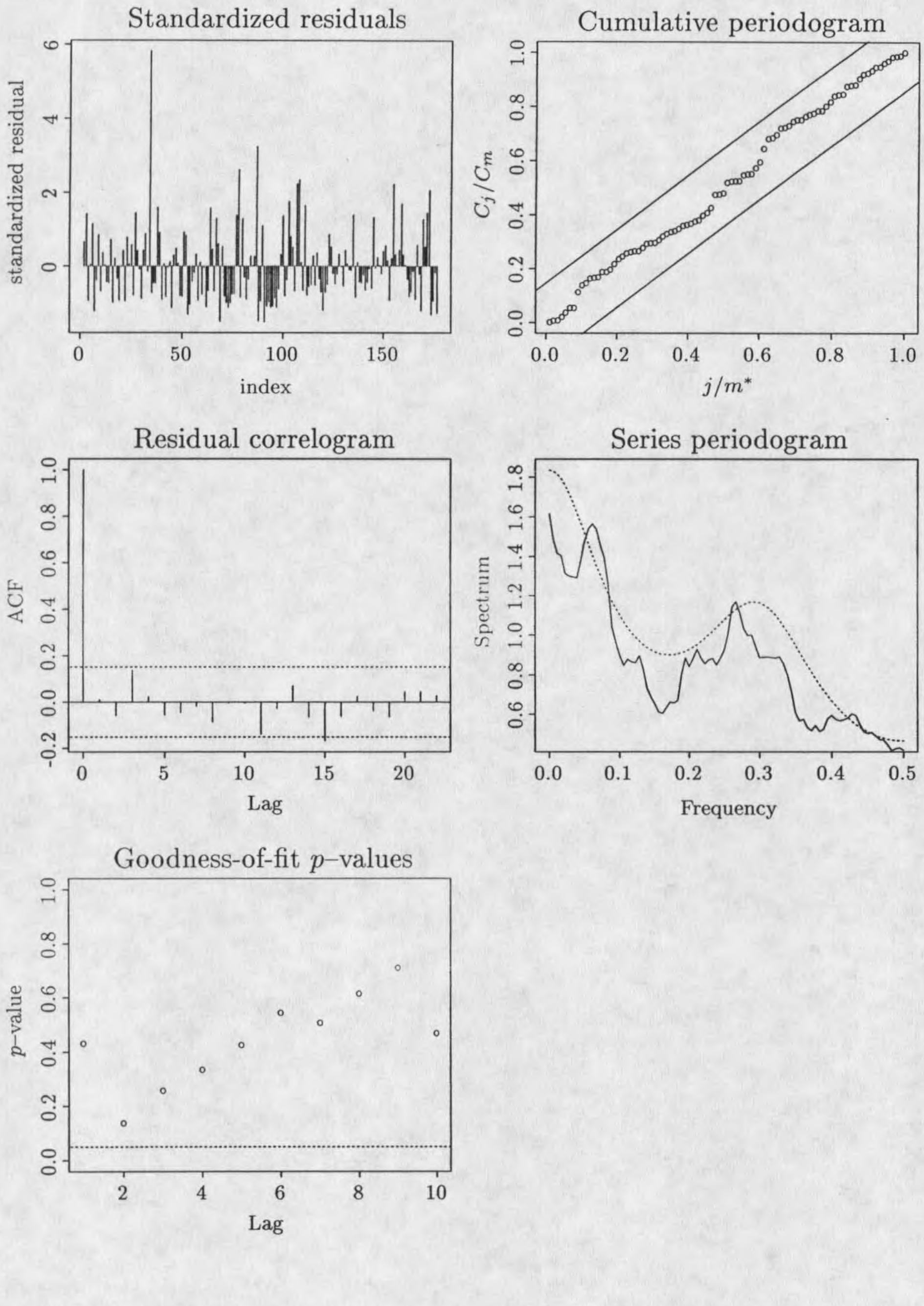

Figure 12: Diagnostic plots for the semi–monthly SNOTEL $AR(1)$ fit.

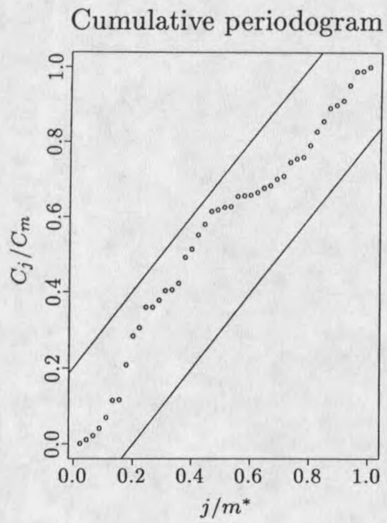

Figure 13: Cumulative periodogram for the monthly SNOTEL white noise model.

spikes and both the serial and seasonal model coefficients test not significant, the model for the monthly SNOTEL data set is pure white noise. It looks like $Y_t = \varepsilon_t$.

Appendix A summarizes model building results for the three SNOTEL series. Estimated coefficients, significance tests, and residual mean squares for each candidate model are included.

Exploratory data analysis shows that substituting the SNOTEL estimated series for the raw series is valid when comparing the SNOTEL and NOAA datasets. ARIMA models for the SNOTEL estimated data sets forego any seasonal coefficients, and, as the averaging period increases, even serial terms get eliminated. $AR(1)$ models adequately represent the weekly and semi–monthly series while the monthly model consists of a white noise term.

Frequency Domain Analysis

Although none of the time domain models include a seasonal component, various exploratory plots support a weak yearly cycle. ARIMA models provide an important tool for forecasting future precipitation, but don't necessarily include a seasonal term when a yearly cycle exists. A major function of frequency domain analysis is to identify cycles in data where other methods may fail. It is capable of isolating a cyclic component completely masked by noise. The SNOTEL data sets seems to fit the description of large noise and weak cycles, so frequency domain analysis is definitely called for.

It begins by reexamining the smoothed periodograms of the SNOTEL series. These are shown without the fitted ARIMA spectra in Figure 14. Dashed lines serve as reference lines and mark frequencies exploratory analysis identified as potentially important: those corresponding to 12–, 6–, 3–, and 1.5–month cycles.

The SNOTEL periodograms clearly indicate the presence of a 12–month cycle in the data; the first peak is squarely centered on the first reference line, which corresponds to a yearly cycle. Other frequencies also seem to be components. In the weekly periodogram peaks appear at frequencies 0.12, 0.16, and 0.2 which identify 2–month, 6–week, and 5–week cycles, respectively. The semi–monthly periodogram shows additional peaks at frequencies 0.10, 0.225, and 0.3, which correspond to approximately 5–, 2.2–, and 1.5–month cycles. The monthly periodogram illustrates a peak at frequency 0.19, or a 5.3–month cycle. Recall that the number of Fourier frequencies for which the periodogram can be calculated depends on sample size. The actual frequencies at which the periodogram is estimated and the effects of smoothing may shift a peak a distance from a component frequency. Taking this into account, the overall component cycles for the SNOTEL series include 12– and 6–month cycles.

Although the ARIMA models for the SNOTEL series don't include a yearly or sub–yearly component, spectral analysis identifies cycles corresponding to these components as significant contributors to the underlying process. These findings reinforce the patterns found in the median traces of Figures 6, 7, and 8. The overall pattern in the data is determined by the differing phases and amplitudes of the two cycles acting in concert. The yearly cycle shows up as higher amounts of precipitation in the winter months and lower amounts during the summer months. The 6–month cycle shows up as peaks in the spring and fall months and troughs during mid–winter and mid–summer.

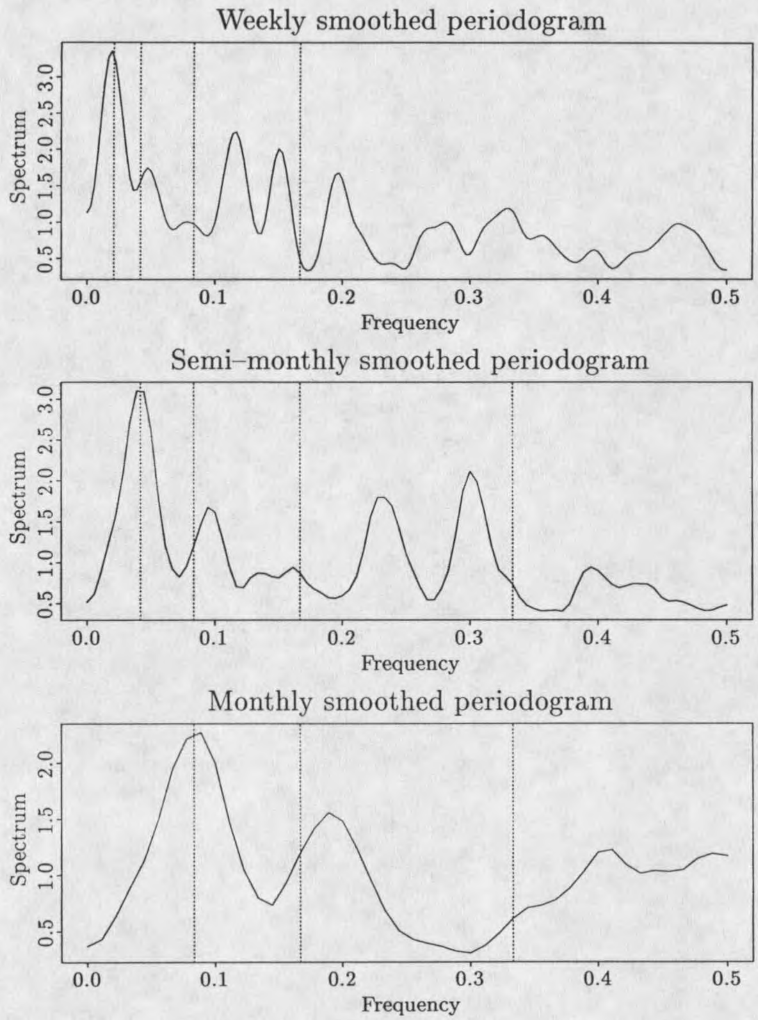

Figure 14: Smoothed periodograms of the SNOTEL estimated data sets. Dashed reference lines identify frequencies with 12–, 6–, 3–, and 1.5–month periods.

Conclusions

The summary statistics and exploratory plots indicate that substituting the estimated series for the raw series is valid for the the SNOTEL data sets. Therefore, the estimated series are used in the ARIMA modeling process. $AR(1)$ models for the weekly and semi–monthly data sets degenerate to a white noise model for the monthly data. The serial correlation breaks down as the averaging period increases.

Although the ARIMA models don't include cyclic terms, frequency domain analysis identifies cycles with 12– and 6–month periods as ones driving the underlying process. The noise component present in the SNOTEL data prevents ARIMA modeling procedures from identifying a yearly component in the any models. As the averaging period increases, the noise component masks any serial or other correlations that might prove significant in a less noisy data set.

CHAPTER 6

NOAA DATA ANALYSIS

Exploratory Data Analysis

The NOAA data set begins January 1, 1969 and ends December 31, 1991. It contains several groups of missing observations, the longest lasting ten months: May 1, 1982 through February 28, 1983. Other periods of missing data range from a single day to two months in length. There are also several observations greater than 1.5 inches, most of them occurring in the 1980s. When the NOAA data is averaged weekly, semi–monthly, and monthly, the extreme observations are absorbed by the averages, except in three cases. In the weekly data set, the extreme observations are the fourth week of December, 1984 with 2.63 inches, the fourth week of February, 1985 with 1.53 inches, and the fourth week of February, 1986 with 1.61 inches. These raw data sets with the missing and extreme observations are compared with corresponding weekly, semi–monthly, and monthly data sets in which missing and extreme observations are estimated and replaced by median values as described in the METHODS section. Table 3 shows summary statistics for the six NOAA data sets.

Figure 15 shows line plots of the NOAA data sets. The raw data sets are in the first column, corresponding estimated data sets in the second. The three extreme observations strongly overshadow the rest of the data, even in the monthly data set. Once these outliers are removed, some interesting features become more noticeable. A slight upward trend of large precipitation averages emerges. That is, the large average values seem to get larger as time progresses. Also, as expected, the large portion

Table 3: Summary statistics for NOAA data sets

	raw series			estimated series		
	weekly	semi–monthly	monthly	weekly	semi–monthly	monthly
mean	0.08	0.08	0.08	0.07	0.07	0.07
st.dev.	0.13	0.09	0.06	0.07	0.05	0.04
min	0.0	0.0	0.0	0.0	0.0	0.0
median	0.05	0.06	0.07	0.05	0.06	0.07
max	0.2.63	0.10	0.52	0.52	0.37	0.24
n	999	502	251	1104	552	276

missing data that has been estimated shows much less variability than surrounding data. Another interesting feature is the strong cyclic pattern appearing from 1987 to roughly 1990. Within this time period, the wet periods consist of several periods in a row with high to moderate precipitation, while the dry periods seem unusually dry. This time period exhibits the strongest yearly pattern seen in either the SNOTEL or the NOAA data sets.

	within series–raw			within series–est		
Averaging period	ANOVA p–value	Test–statistic	Critical value	ANOVA p–value	Test–statistic	Critical value
weekly	0.193	16	2.3	0.529	2.25	2.3
semi–monthly	0.217	10.24	2.3	0.422	2.25	2.3
monthly	0.211	13.44	2.3	0.757	1.78	2.3

	Between series		
Averaging period	ANOVA p–value	Test–statistic	Critical value
weekly	0.073	16	2.3
semi–monthly	0.103	16	2.3
monthly	0.171	13.44	2.3

Table 4: ANOVA p–values for differences between sub–series means and Cochran's test–statistics and critical values for homogeneity of variance between sub–series illustrated in Figure 16.

Figure 16 shows each averaged data set divided into four sub–series of nearly equal length so that stationarity conditions can be checked. The means and standard

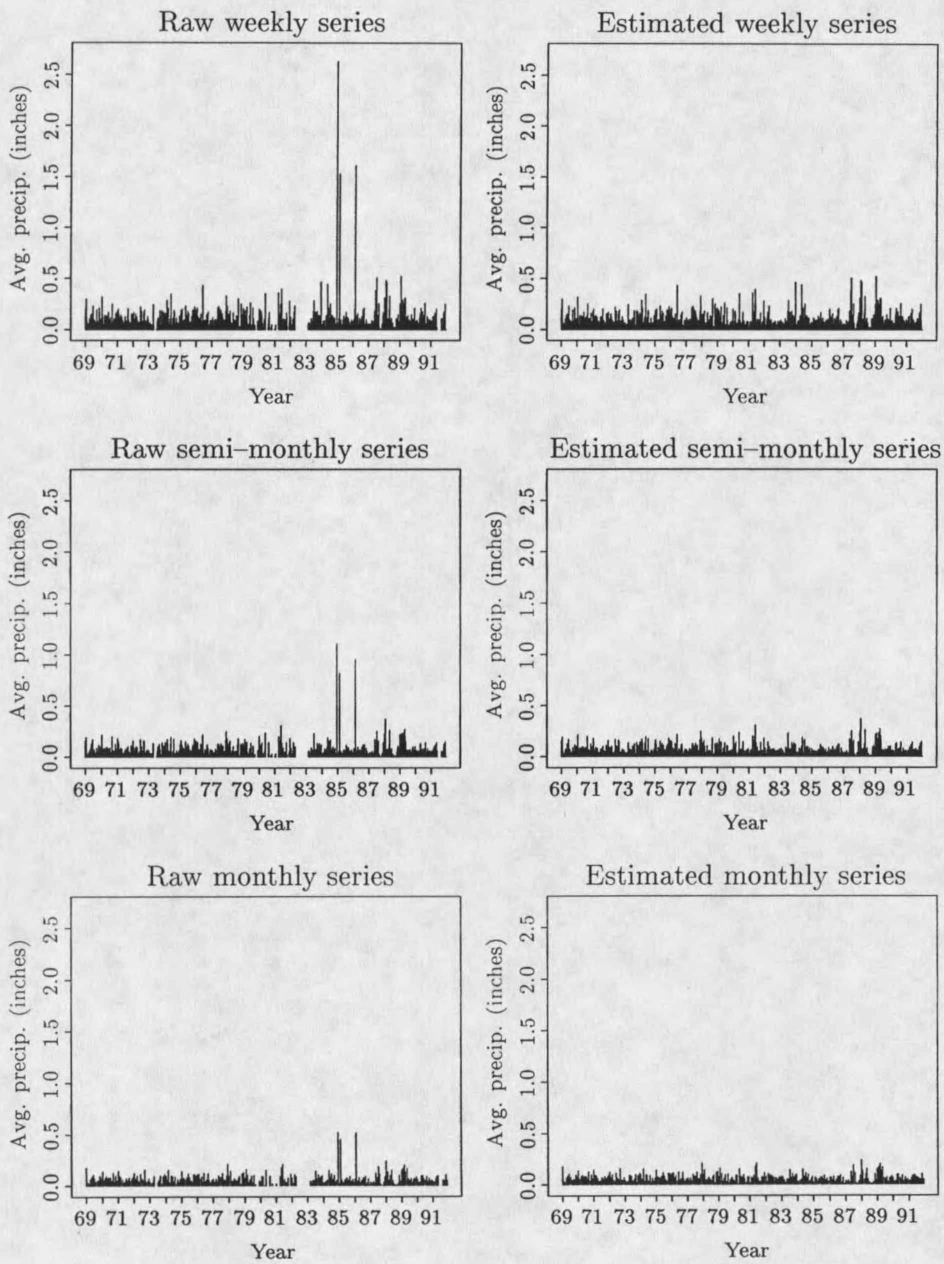

Figure 15: Line plots of the NOAA raw and estimated data sets for weekly, semi–monthly, and monthly averaging periods.

deviations of the sub–series may be tested for significant differences. If any are found, the NOAA data might not be stationary. All three extreme values and the largest portion of missing data fall into the same sub–series. As expected, once the outliers are removed and the less variable estimates are included, the means and standard deviations of these sub–series decrease for every averaged series. The weekly mean is the most affected mean, while the standard deviation shows a marked decrease in each data set. If any significant differences are found, they will most likely be here. The ANOVA p–values for the means and Cochran's test–statistics and .01 critical values appear in Table .

The ANOVA test for $H_o : \mu_1 = \mu_2 = \mu_3 = \mu_4$ returns no significant p–values for any of the series and Cochran's critical values are much smaller than any of the test–statistics. Therefore, second order stationarity may be assumed for both the raw and estimated series. Studying the effects of estimating outliers and missing observations can be studied through continuing exploratory analysis.

Lowess smooths of the NOAA data sets are shown in Figure 17. The first column of plots shows the raw data sets, excluding the outliers. The smooths are calculated with the outliers present, but were left out of the plots for scale reasons. The second column shows plots of the estimated data and corresponding smooths. The missing values in the raw data set were replaced with the median of the entire series in the calculation of the raw smooths because the smooth function does not handle missing data. Because of this, the raw smooths are quite flat throughout that period. The raw and estimated smooths are nearly identical–the outliers have little effect on the raw smooth because of the robust nature of the smoothing function.

Figure 18 displays line plots of the residuals from the mean, with the raw residuals in the first column and estimated residuals in the second. Because of the large amount of data, the smooths in the weekly plots are difficult to see. Therefore,

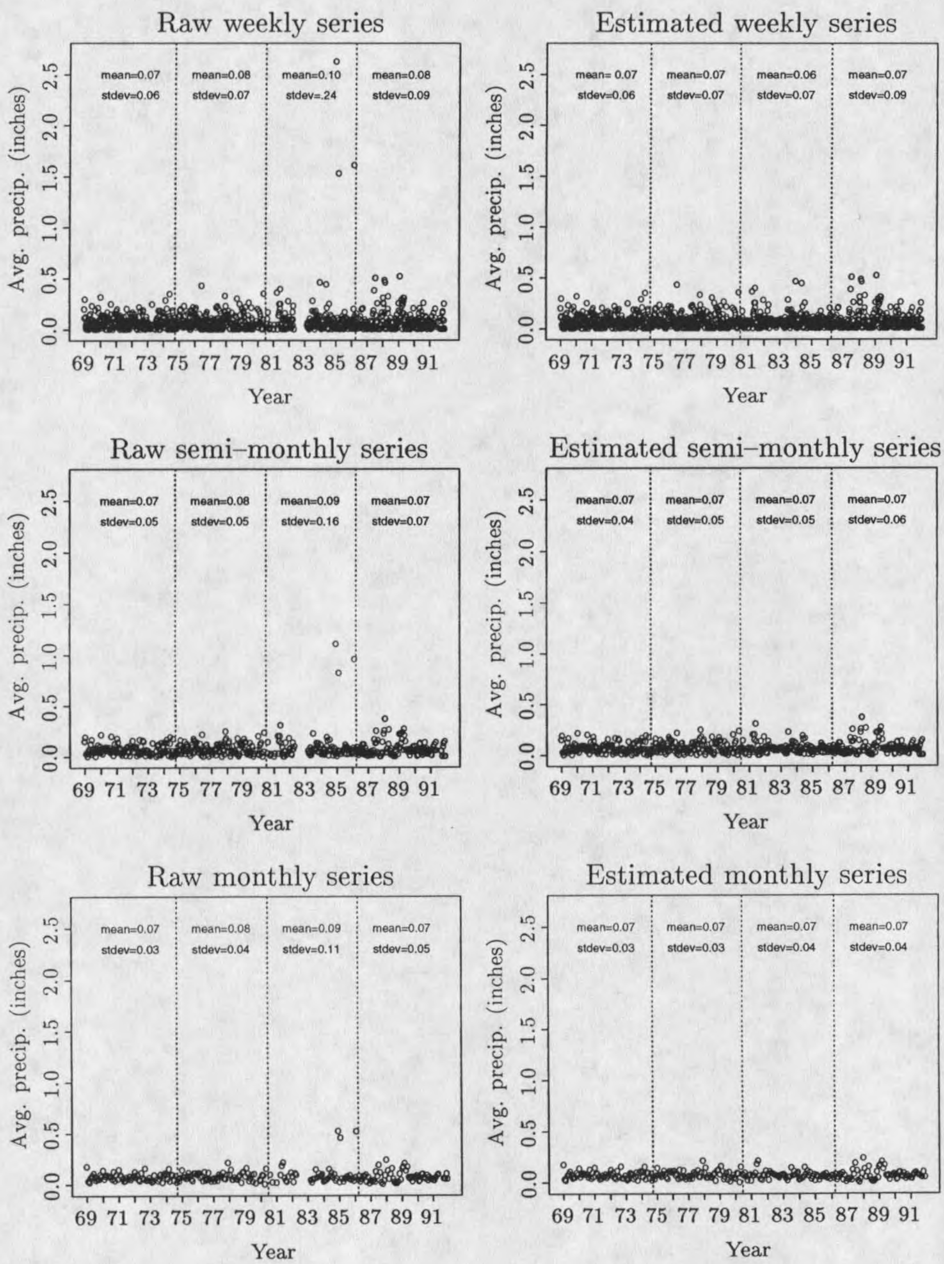

Figure 16: NOAA raw and estimated weekly, semi–monthly, and monthly data sets are divided into fourths. Means and standard deviations for each series are included.

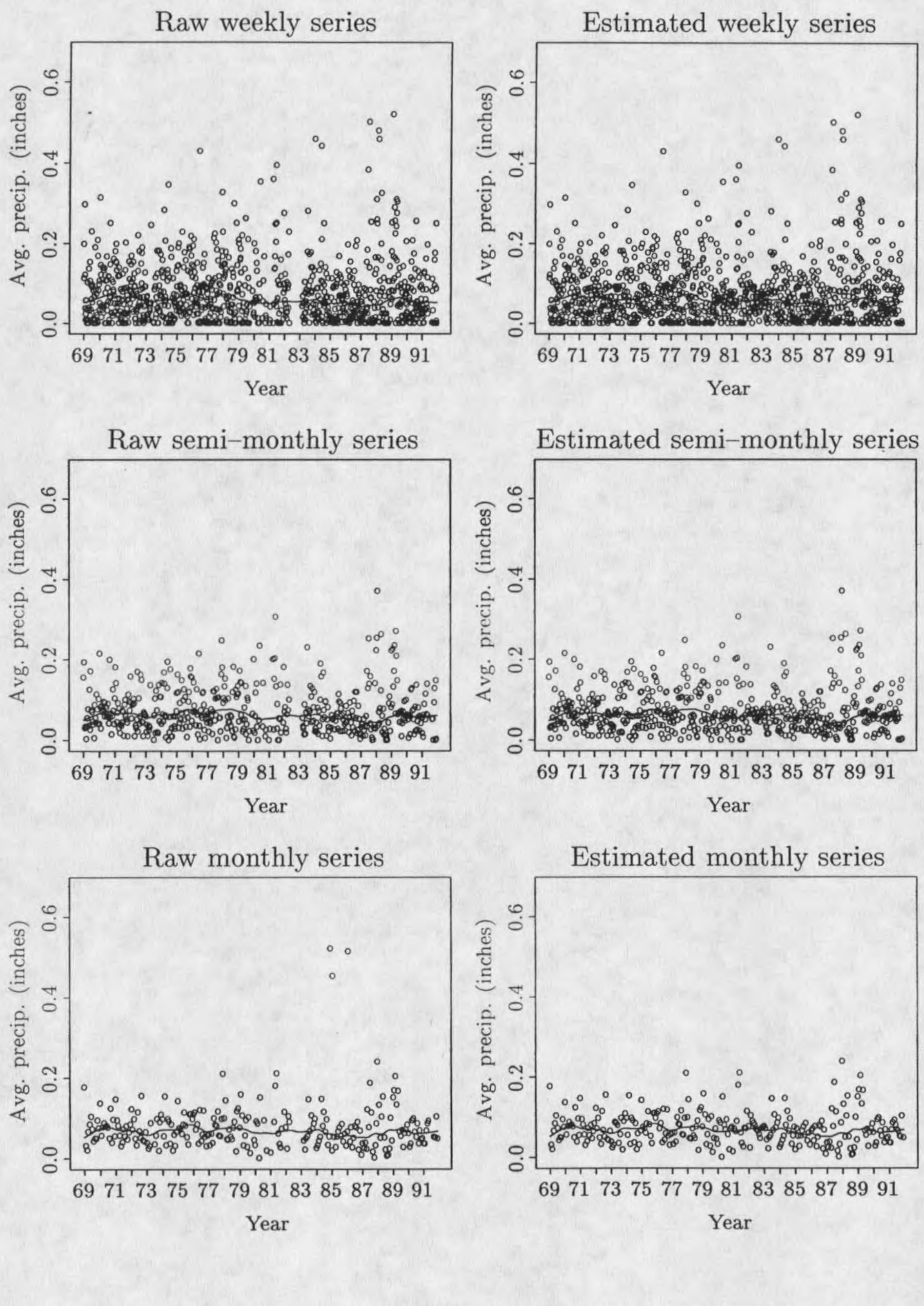

Figure 17: Lowess smooths of raw and estimated NOAA series with smoothing parameter $f = .1$.

potential cyclic behavior and differences between raw and estimated data are difficult to distinguish. The semi–monthly and monthly smooths are more interesting to look at. In this series of plots, zero replaces the missing residuals in the calculation of the smooth, causing a flat smooth in the large areas of missing data. The smooths from the raw and estimated data sets follow nearly identical patterns. As the averaged time increases, a cyclic pattern becomes evident, with peaks occurring approximately eighteen months apart.

Figure 19 shows the NOAA data plotted by year. These plots are not particularly useful for showing any yearly cycle in the data because of the data's noise. The only noticeable difference between the raw and estimated data is the absence of outliers. Otherwise, the plots telegraph little information.

Boxplots, median traces, and 95% pointwise confidence intervals for the medians present more information about the yearly cyclic behavior of the data and are shown in Figures 20, 21, and 22. Weekly, semi–monthly, and monthly plots for the raw and estimated data are shown in 20, 21, and 22 respectively. Raw data plots are shown in the first column, their estimated counterparts in the second. Similarities and differences between the raw and estimated data are discussed before the general behavior of the series is examined across time intervals.

The boxplots are nearly identical from raw to estimated data sets. The only subtle difference between the two is the length of some whiskers. A few of the raw boxplots have longer whiskers due to outliers. The lengths of the boxplots themselves also differ slightly because of the differing lengths of the raw and estimated data sets. For example, if the first week of October, 1985 is missing in the weekly data set, the raw data has 29 observations for that period while the corresponding estimated data set has 30 observations to calculate the interquartile range for that period.

The median traces of the raw and estimated data sets look identical across all

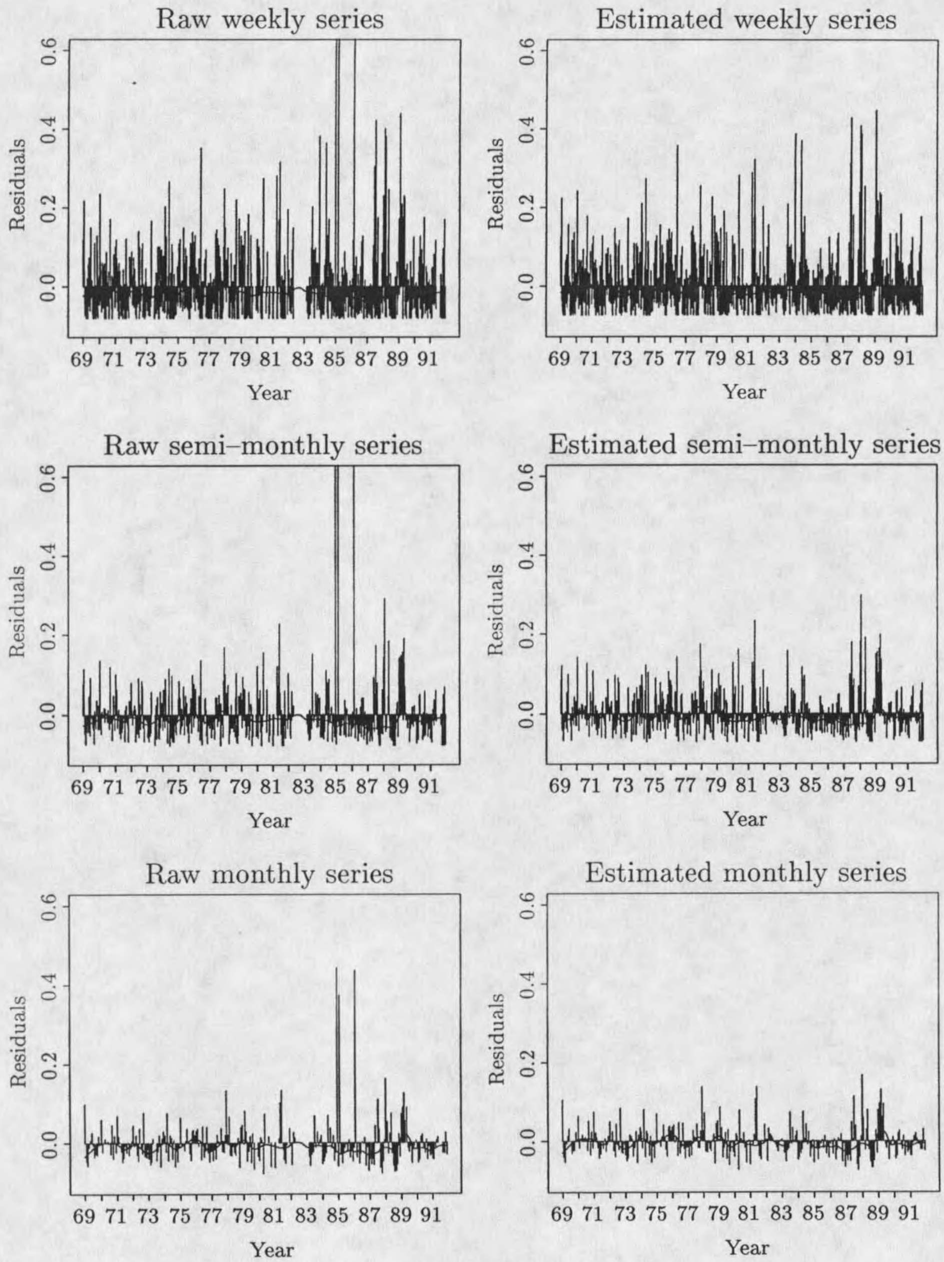

Figure 18: NOAA residuals from the mean for weekly, semi–monthly, and monthly data sets. The lowess smoothing parameter is $f = .06$.

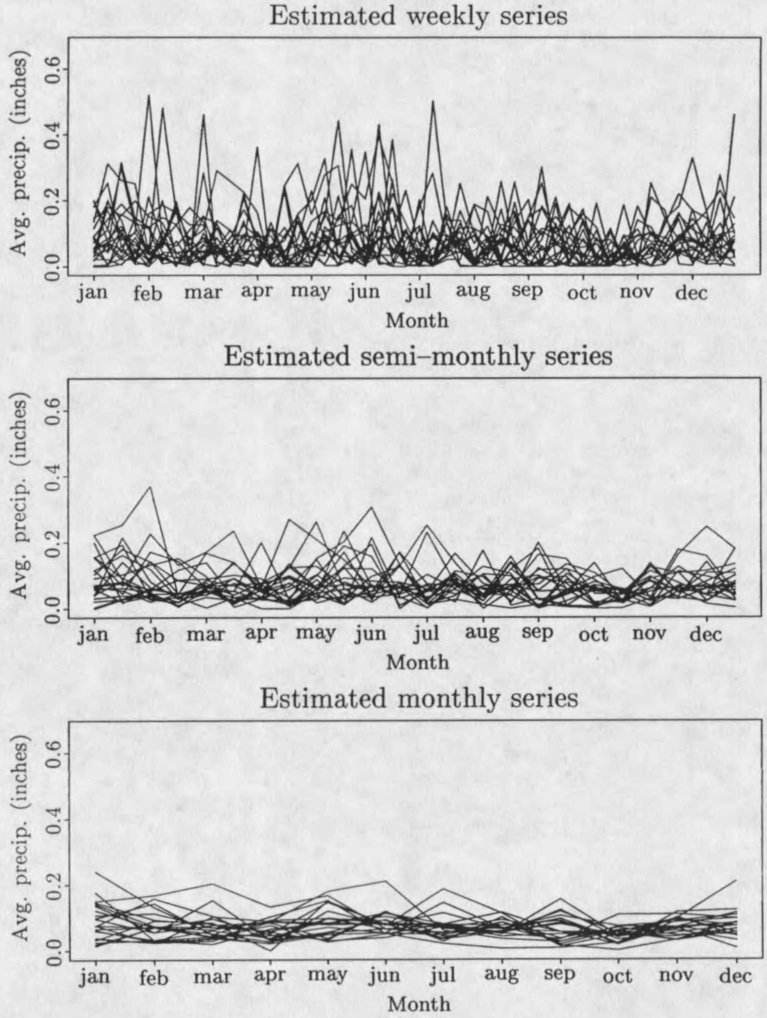

Figure 19: NOAA weekly, semi–monthly, and monthly estimated data sets. Each line represents a year of average precipitation values.

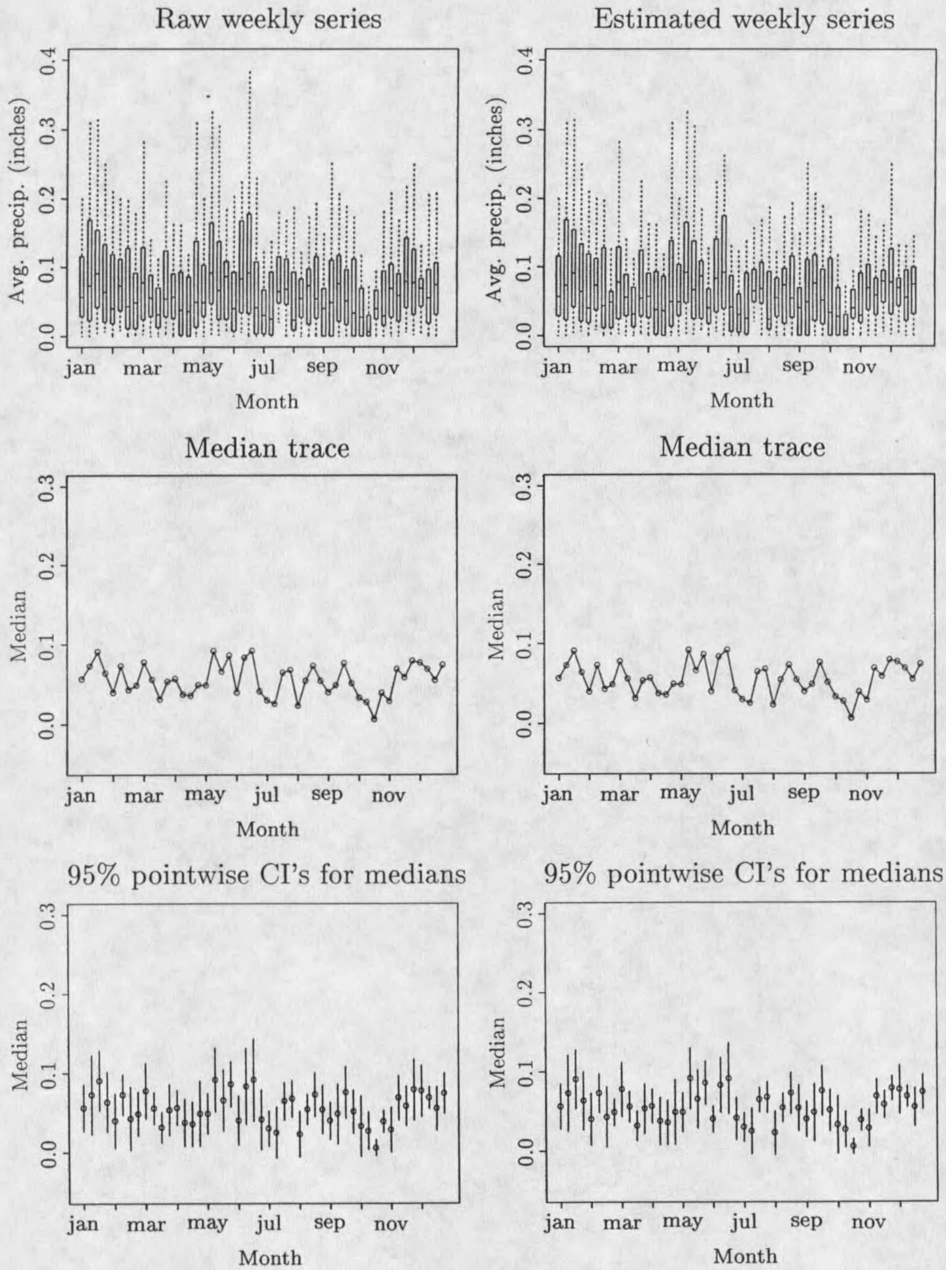

Figure 20: Boxplot, median trace, and 95% confidence intervals for the medians for NOAA weekly raw and estimated series. Outliers are omitted.

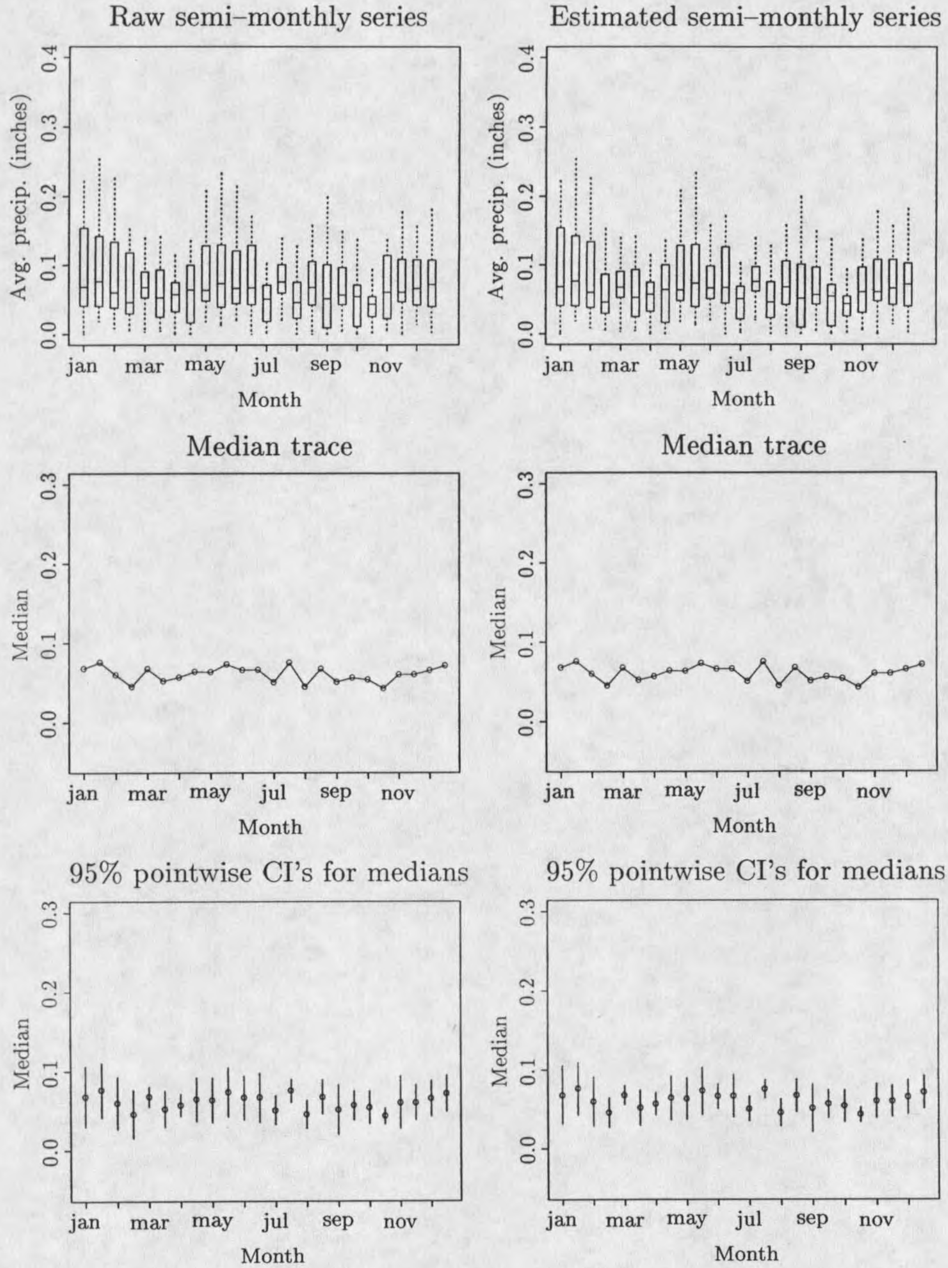

Figure 21: Boxplot, median trace, and 95% confidence intervals for the medians for NOAA semi–monthly raw and estimated series. Outliers are omitted.

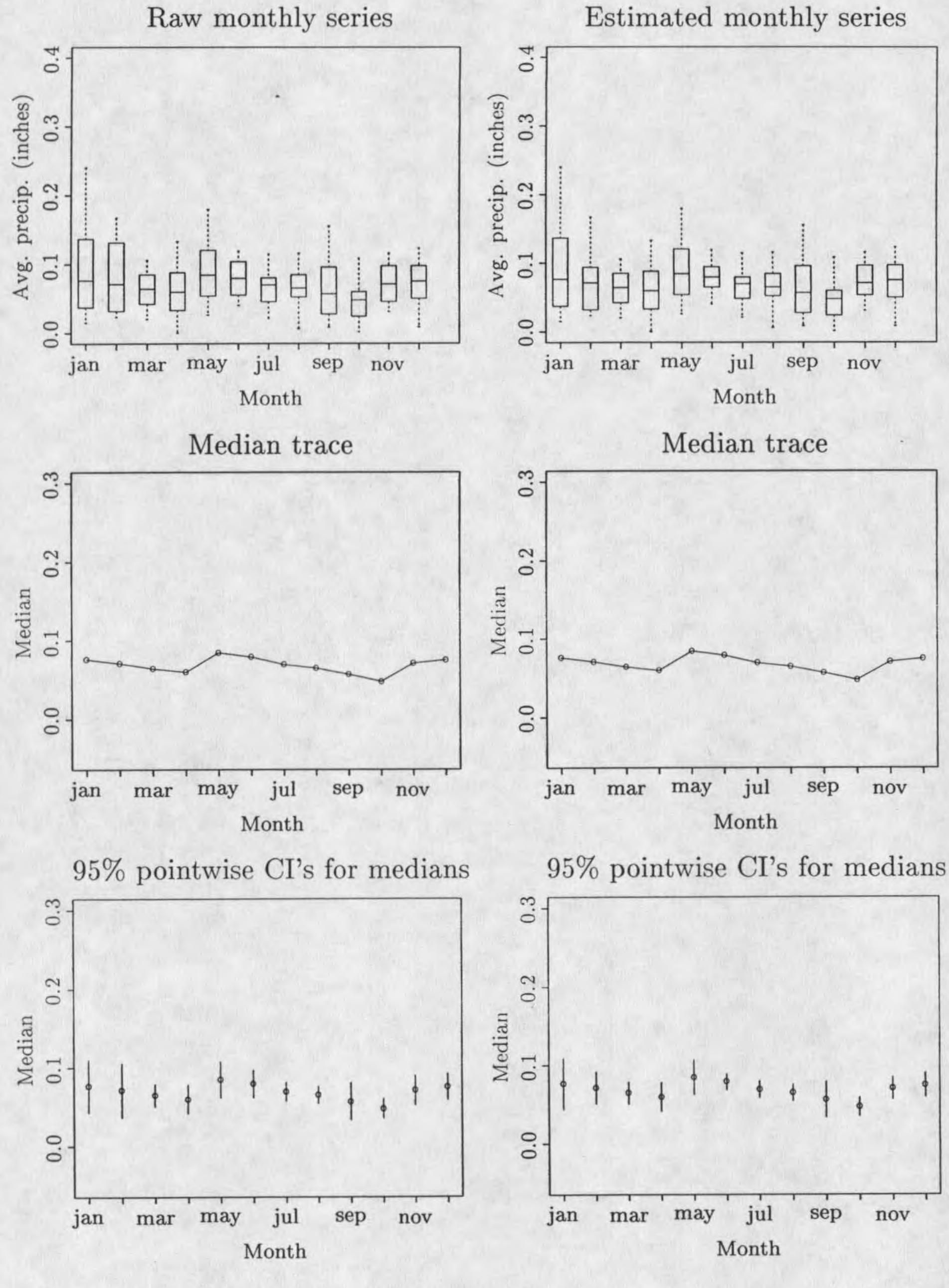

Figure 22: Boxplot, median trace, and 95% confidence intervals for the medians for NOAA monthly raw and estimated series. Outliers are omitted.

averaged periods. This identity is expected for two reasons. The first is that outliers do not affect the median, so the raw median traces do not jump up in the periods where the outliers are located. The second reason is that missing data also does not affect the location of the median. Because the missing data has been estimated by the medians of the present data, the medians of the estimated data match those of the raw data. The widths of the 95% pointwise confidence intervals for the medians differ slightly between raw and estimated because of the larger number of observations in some of the estimated periods. As with the interquartile ranges, the missing data are not included in the calculation of confidence intervals. The larger number of observations in the estimated data set results in a narrower confidence interval for some weeks and months.

The general behavior of the series across time intervals can be addressed now that differences caused by the differing lengths of the time series have been discussed. First, the skewed nature of the NOAA data is shown by the longer, top whiskers of the boxplots. In general, however, the medians fall near the middle of the boxplots. Although less noticeable in the weekly boxplots, the narrowest boxplots occur in the summer months and the widest in the winter months. The boxplots show no yearly cyclic pattern of wet spring and late fall and dry winter and summer behavior, as expected.

Second, the median traces proved a summary of yearly precipitation by averaging period. The weekly medians exhibit a high frequency cycle, perhaps six weeks in length. The semi–monthly medians remain fairly constant, making identification of any cycle by eye alone difficult. The monthly medians seem to show a weak yearly pattern of decreasing precipitation from spring to summer before increasing in mid–to–late fall.

Third, the 95% pointwise confidence intervals for the medians reiterate the patterns, or lack thereof, found in the median traces. The weekly confidence intervals show random fluctuation; the semi–monthly intervals remain fairly level and their widths remain relatively constant; the monthly intervals show a weak cyclic pattern similar to the one described above.

Time Domain Analysis

Thus far, tables and plots summarizing the NOAA data set helped identify cycles that may be important. Plots of the autocorrelations and partial autocorrelations might uncover other cycles previously undetected. Figures 23 and 24 show the correlograms and partial correlograms of the raw data in the first column and estimated data in the second column of each figure. Because the Splus function that calculates the ordinates does not handle missing data, correlograms and partial correlograms for the raw data are calculated with medians replacing the missing values.

Figure 23 shows correlograms for each NOAA series. Each of the raw correlograms has positive peaks at lags equivalent to 2, 12, and 14 months, indicating positive 2–month and yearly correlations between observations in the raw data set. At all other lags, even lag 1, the spikes are insignificant, signaling no correlation between observations at those lags. The notion of no serial correlation between observations, even in the weekly data, is surprising. The correlograms of the raw data indicate a purely seasonal model, which is somewhat unrealistic based on the general knowledge that weather follows an autoregressive process.

The correlograms of the estimated data are more realistic and better behaved. They are shown in the second column of Figure 23. Overall, the spikes are less significant than the raw correlogram spikes but exhibit expected results. The weekly correlogram oscillates, indicating positive and negative correlation between observations over the seasons. It starts out with positive lags that decrease in significance. At

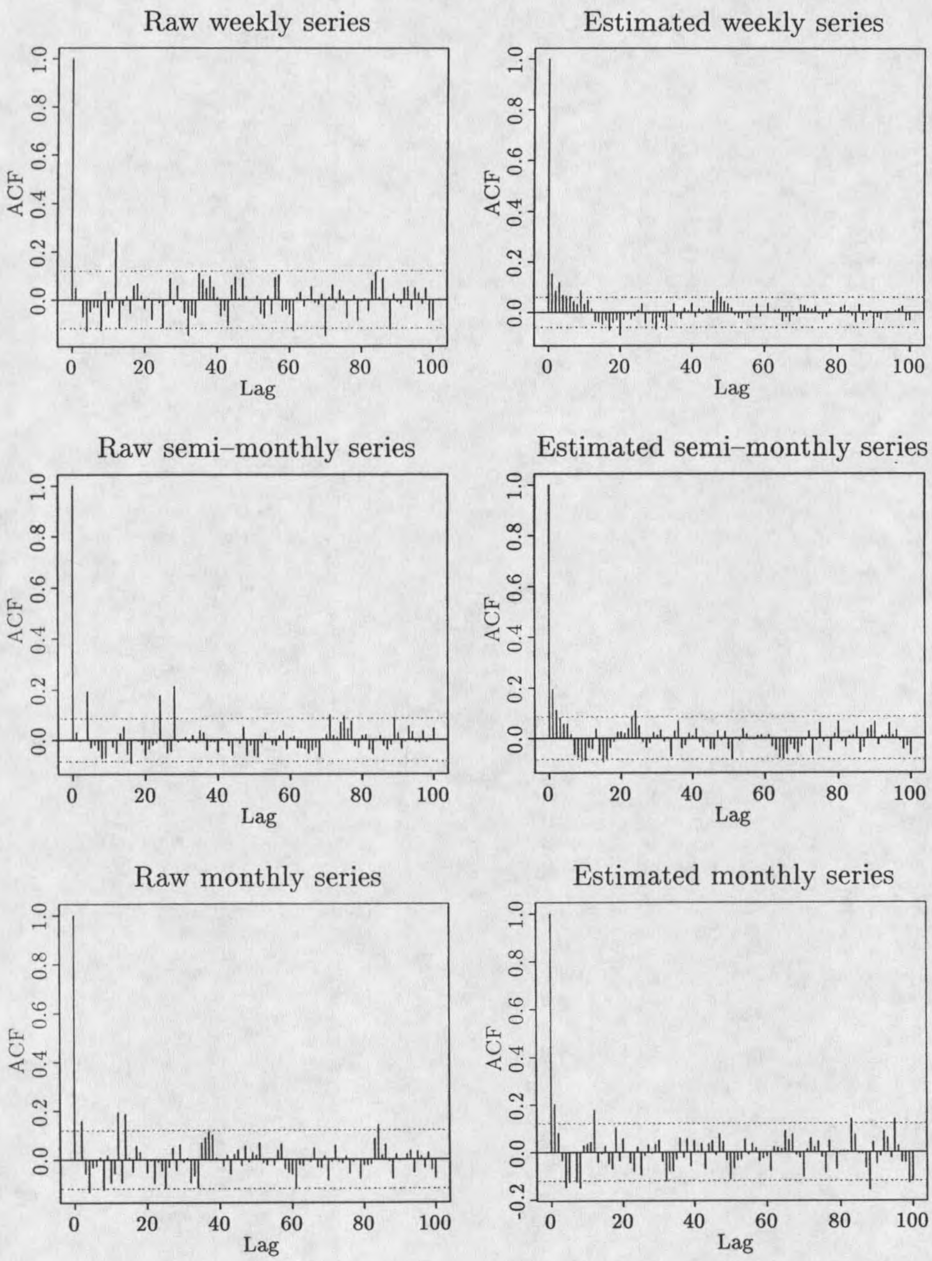

Figure 23: NOAA sample autocorrelation functions for weekly, semi–monthly, and monthly data sets.

the three month mark, the spikes become negative with small significance, indicating slight negative correlation between observations at those lags. As the lags approach a year in length, the spikes again become positive, with a significant spike at exactly a year. The correlograms for the estimated semi–monthly and monthly data sets show similar results. Each shows positive serial and yearly correlation with negative correlations in the mid–year lags.

The partial correlograms are shown in Figure 24. Once again, the first column shows partial autocorrelations in the raw data where the missing values have been replaced by median values, and the second column shows the same for the estimated data sets. The partial correlograms reflect the relationships in the data found in the correlograms by exhibiting similar behavior. Partial autocorrelations at lags 2, 12, and 14 months are indicated by positive spikes at those lags. The raw plots also show significant negative spikes at lags equivalent to 4 months, indicating negative correlation between observations 4 months apart once the effects of other correlations have been removed. The cyclic nature of the correlations is more pronounced in the estimated partial correlograms, which parallels the correlogram behavior. The significance of individual partial autocorrelations has decreased, but the cyclic behavior is more prominent. Spikes for small lags are positive, mid–year lags have negative ordinates, and lags near a year in length are once again positive, indicating a yearly cycle of correlations among observations.

Patterns in the correlograms and partial correlograms suggest AR or ARMA models as candidates for modeling the NOAA series. Regular and seasonal models may be appropriate for differing data sets. ARMA(1,1) models provide another possibility. The similarities in the raw and estimated data sets shown throughout the exploratory data analysis provides evidence that the estimated data may be substituted for the raw when comparing the NOAA and SNOTEL series. Trial overfitting

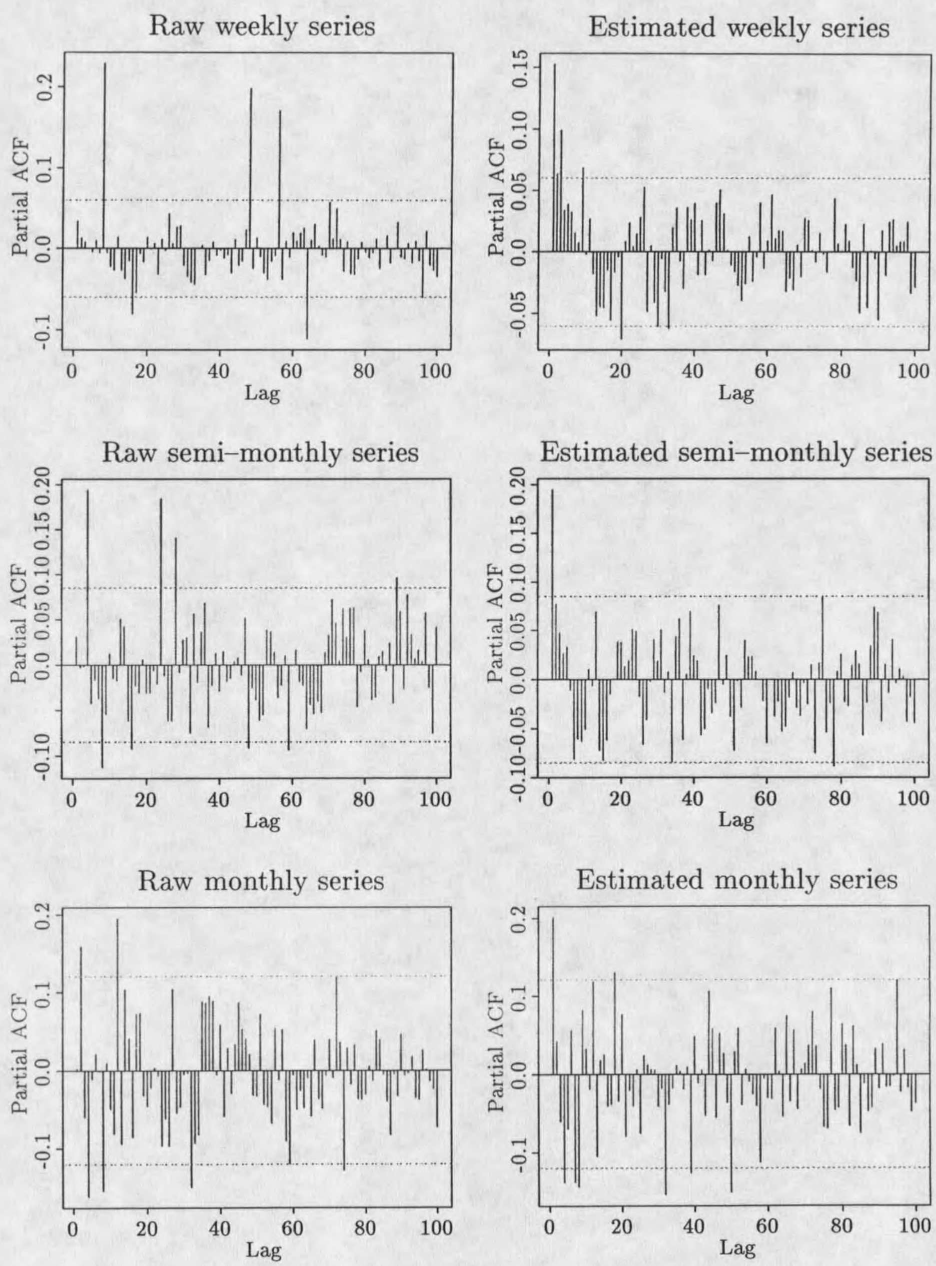

Figure 24: NOAA sample partial autocorrelation functions for weekly, semi–monthly, and monthly data sets.

combined with diagnostic plots lead to a model for each of estimated NOAA series.

Because of the small lag correlations exhibited in the weekly correlogram and partial correlogram, an AR(3) is the first model selected. Estimated coefficients are $\hat{a}_1 = .1351$, $\hat{a}_2 = .1453$, $\hat{a}_3 = .0999$. The residual mean square is 33.18 and a t–test for $H_o : a_3 = 0$ is .0009. Diagnostic plots for the AR(3) model are shown in Figure 25.

The standardized residuals, residual correlogram, and cumulative periodogram indicate the residuals are white noise. The goodness of fit plot shows no significant evidence of correlated residuals at lags 1 through 10. The raw periodogram of the data follows the patterns in the theoretical AR(3) spectrum for the fitted model. Diagnostics indicate the AR(3) model is appropriate.

Trial overfits of the model include an AR(4), ARMA(1,1), and a seasonal AR(1)(1)48. Residual mean square for each of the models is 33.16, 33.12, and 32.48, respectively. Although the AR(1)(1)48 model has a somewhat lower residual mean square than the AR(3) candidate, the p–value for $H_o : \tau_1 = 0$ is .267. Because the overfit models only slightly reduce the residual mean square and the AR(3) model passes diagnostic checks, it makes up the final model. It looks like $Y_t = .1351Y_{t-1} + .0453Y_{t-2} + .0999Y_{t-3} + \varepsilon_t, \quad t = 4, 5, \ldots, 1104$.

The correlogram of the NOAA semi–monthly series shows serial autocorrelations as well as slightly significant yearly autocorrelation. Candidate models include AR(1) and AR(1)(1)24. AR(2), AR(2)(1)24, and ARMA(1,1) are possible overfits. The AR(1) fit returns a residual mean square of 23.473 and a p–value of 3.10×10^{-6} for the first coefficient. Obviously the final model needs the a_1 coefficient. The AR(1)(1)24 fit returns the coefficients $\hat{a}_1 = .1925$ and $\hat{\tau}_1 = .0953$ and a seasonal p–value of .028. The residual mean square for the seasonal model is 22.956. The diagnostic plots for the seasonal model appear in Figure 26. The standardized resid-

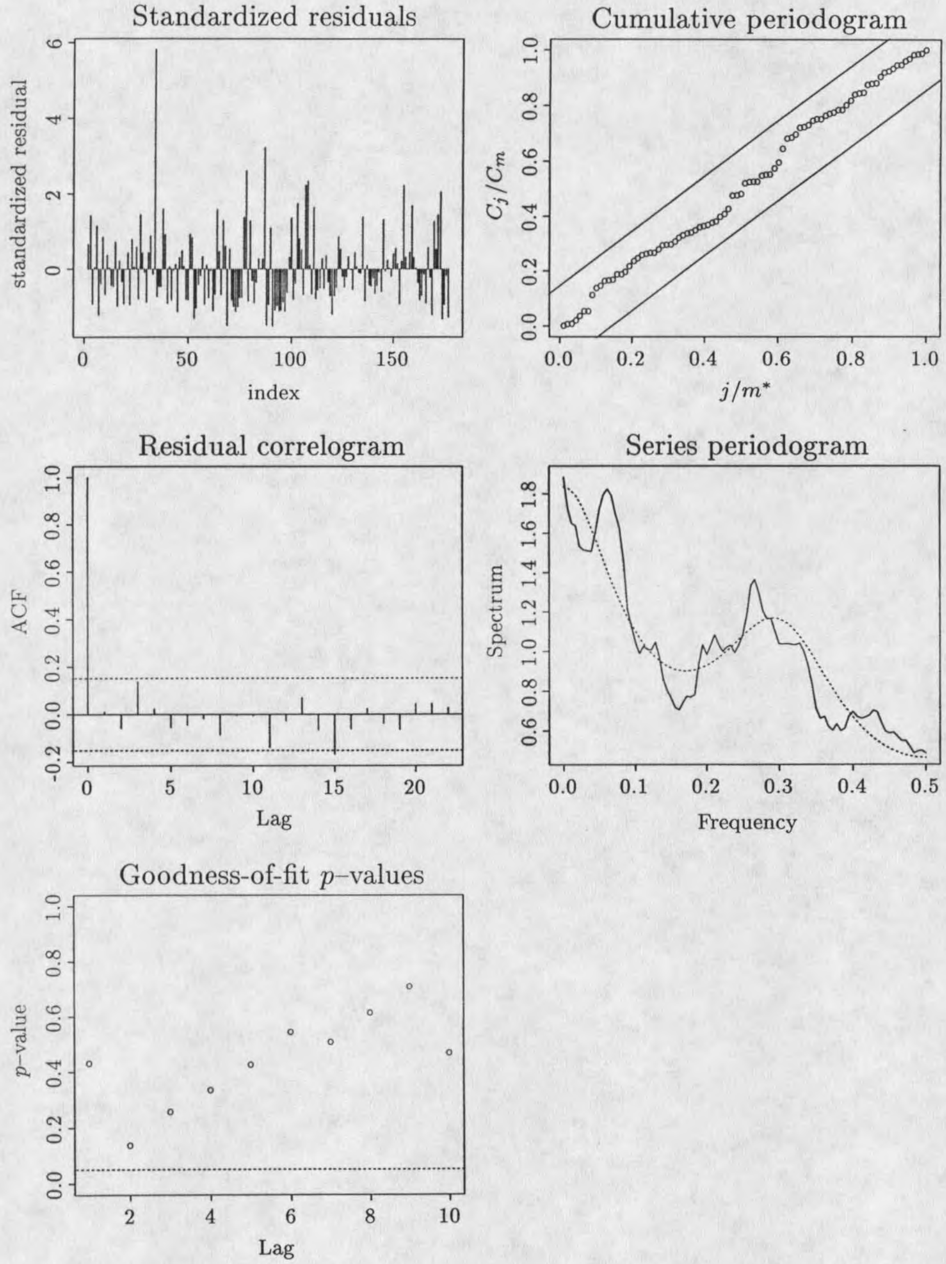

Figure 25: Diagnostic plots for the weekly NOAA $AR(3)$ fit

uals, residual correlogram, and the cumulative periodogram indicate the residuals are a white noise sequence. The large p–values in the goodness of fit plot also provide evidence of white noise residuals. The theoretical spectrum and smoothed periodogram do not behave as similarly as those of some previous series but the cyclic patterns of the spectrum are repeated in the smoothed periodogram.

The AR(2), AR(2)(1)24, and ARMA(1,1) overfits return residual mean squares of 23.45, 22.93, and 23.47, respectively. None of the additional coefficients tested significant. The diagnostic plots and overfits support the AR(1)(1)24 model of the form $Y_t = .1925Y_{t-1} + .0953Y_{t-24} + \varepsilon_t, \quad t = 25, 26, \ldots, 552.$

The correlogram for the NOAA monthly series shows significant peaks at lags 1 and 12 while the partial correlogram shows significant peaks at lags 1, 4, 7, 8, and 32. The differences in the two makes listing candidate models somewhat difficult. Because lags in multiples of 4 are significant in the partial correlogram, one possible solution would be to difference the data by lag 4 and model the resulting series as an AR(1), AR(2), or AR(2)(1)8. These models correspond with the spikes at lags 4, 8, and 32.

It's best to keep the model as simple as possible by diagnosing models that forego differencing. If those models prove inadequate, a more difficult model that includes differencing might then be considered. The differenced models would not include the serial correlation shown in both the correlogram and partial correlogram. If AR(1) or AR(1)(1)12 prove unsatisfactory, differencing provides an alternative. An AR(1) fit returns a coefficient $\hat{a}_1 = .2013$ and a residual mean square of 16.58. The p–value for $H_o : a_1 = 0$ is .0007, indicating the serial correlation is significant. The $AR(1)(1)12$ fit returns coefficients of $\hat{a}_1 = .2270$ and $\hat{\tau}_1 = .1861$. The residual mean square for the seasonal model is 16.217 and the p–value for the seasonal coefficient is .0024. The diagnostic plots for the seasonal model appear in Figure 27.

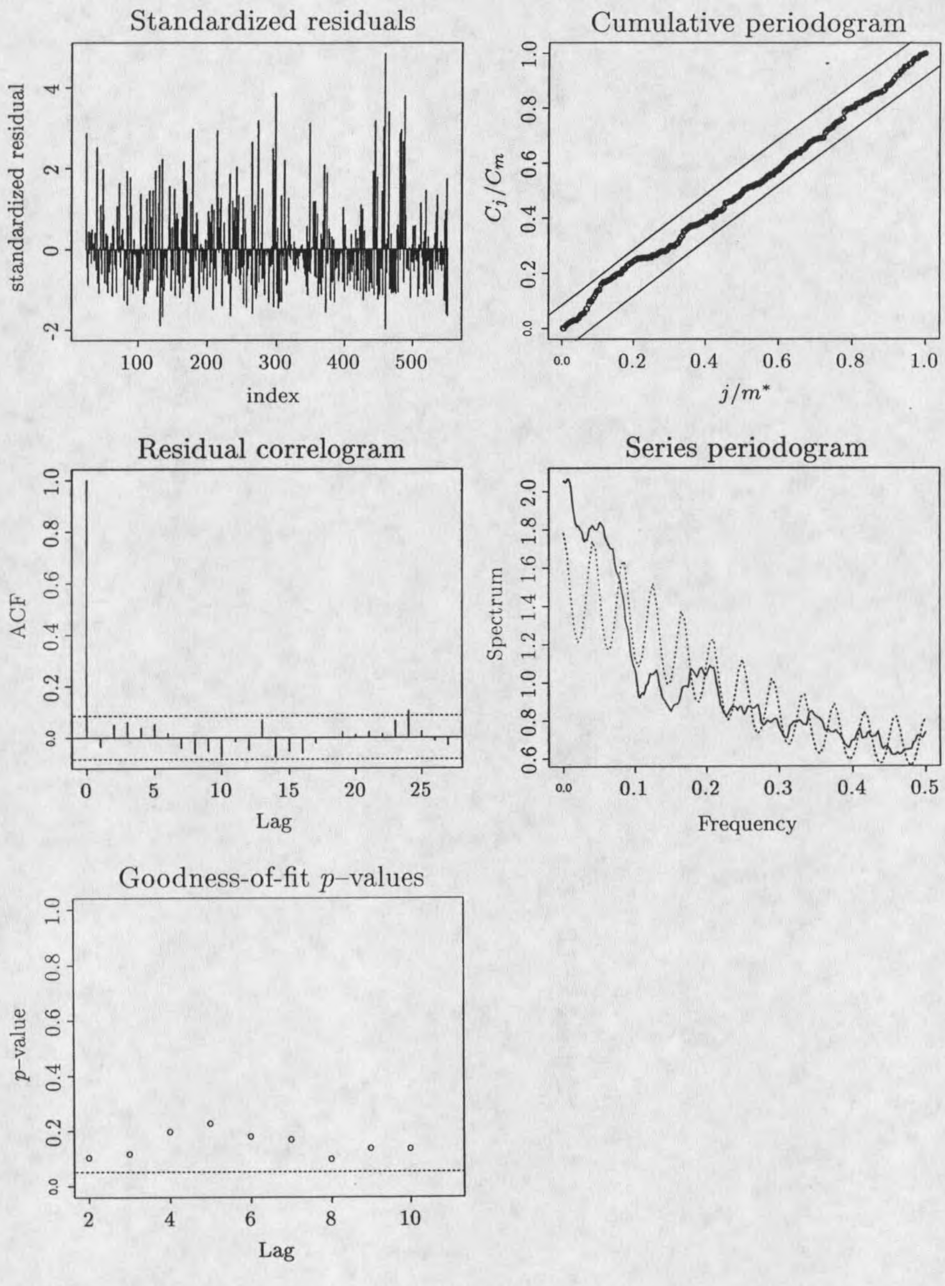

Figure 26: Diagnostic plots for the semi–monthly NOAA $AR(1)(1)24$ fit

The standardized residuals, residual correlogram, and cumulative periodogram of the residuals support a white noise residual sequence. The significant p–values in the goodness–of–fit statistic correspond with spikes at lags 7 and 8 in the partial correlogram. Trial overfits do not significantly reduce the residual mean square and none of the additional coefficients test significant. Although the goodness–of–fit plot shows correlation among the residuals at lags 7 and 8, the other diagnostic plots indicate the seasonal model adequately reduces the residuals to white noise. Because a major goal of the research is to search for yearly patterns and the seasonal model meets this objective, it is adequate. The final model for the monthly NOAA series is $Y_t = .2270Y_{t-1} + .1861Y_{t-12} + \varepsilon_t$.

Appendix A summarizes model building results for the three NOAA series. Estimated coefficients, significance tests, and residual mean squares for each candidate model are included.

Exploratory data analysis on the NOAA data sets shows that substituting the estimated series for the raw series is valid when comparing the NOAA and SNOTEL data sets. The natural smoothing that takes place when considering means is illustrated by the change in ARIMA models as the averaging period increases. The weekly model, an AR(3), takes advantage of small lag correlations in the data. The autocorrelations at lags 2 and 3 get absorbed into the semi–monthly and monthly observations, resulting in AR(1) regular components in the semi–monthly and monthly models. As the averaging period increases, seasonal coefficients become more important. Seasonal correlations appear weaker in the weekly data because of the strong small lag autocorrelations. In the semi–monthly and monthly series, however, seasonal terms significantly improve the models. Results of a frequency domain analysis on the NOAA data should agree with the semi–monthly and monthly models by uncovering yearly or sub–yearly cycles. It may even be able to identify such cycles in

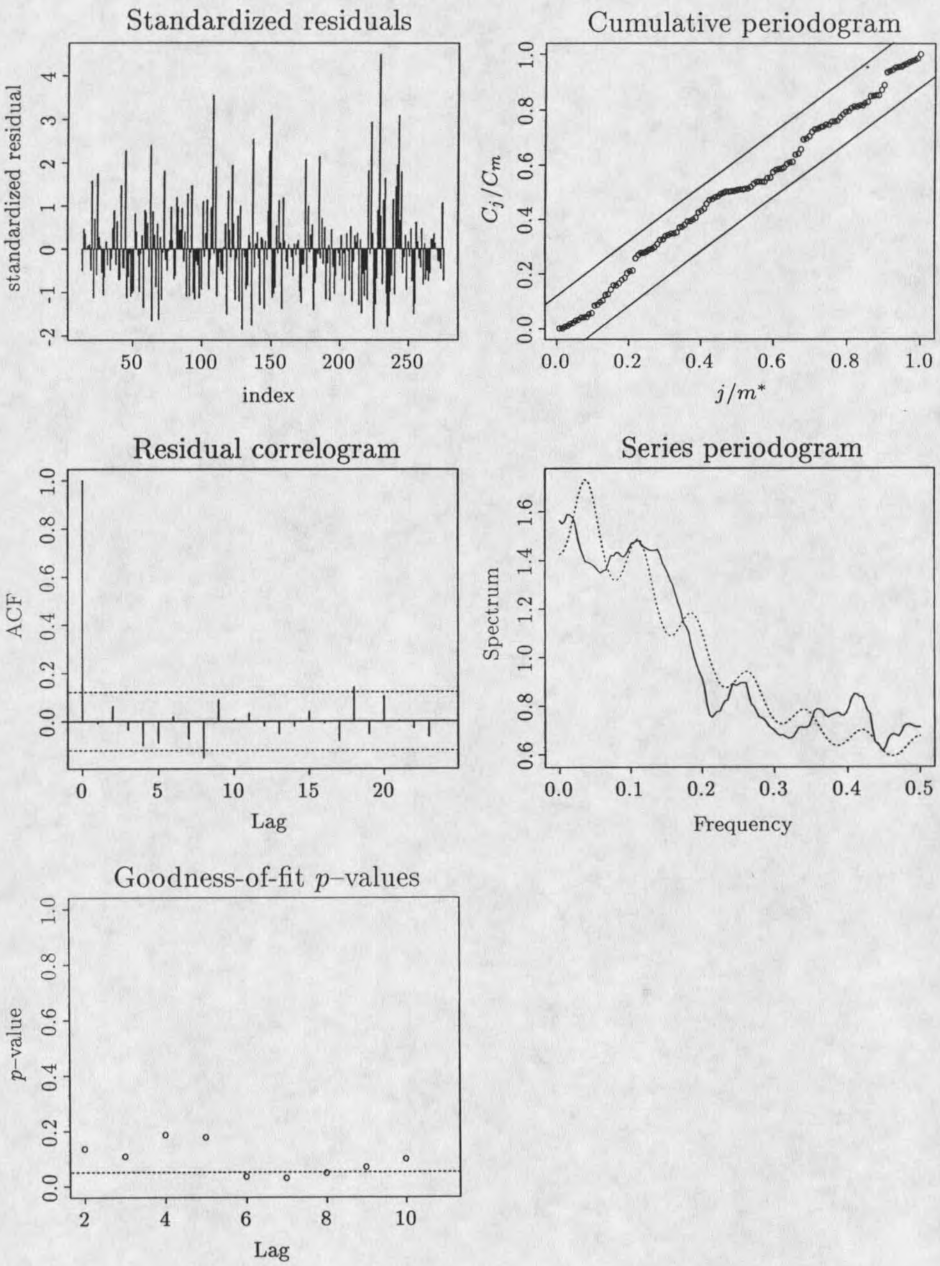

Figure 27: Diagnostic plots for the monthly NOAA $AR(1)(1)12$ fit

the weekly data, even though ARIMA modeling procedures didn't.

Frequency Domain Analysis

Frequency domain analysis begins by examining the smoothed periodograms of the NOAA estimated data sets which appear in Figure 28. Once again, dashed reference lines mark the frequencies associated with 12-, 6–, 3–, and 1.5–month cycles.

As in the SNOTEL periodograms, evidence of a yearly cycle is shown by the first peak in all three periodograms squarely centered over the yearly cycle reference line. The yearly frequency dominates in the NOAA series, but other cycles also contribute to the variation in the series. The weekly periodogram peaks at frequencies 0.08 and 0.32 which correspond to approximately 3– and 1–month cycles, respectively. The semi–monthly periodogram shows additional peaks at frequencies 0.08 and 0.167, corresponding to 6– and 3–month cycles. Finally, the monthly periodogram shows peaks at 0.167 and 0.33, which also correspond to 6– and 3–month cycles. Clearly 12–, 6–, and 3–month cycles contribute significantly to the observed signal.

Conclusions

The summary statistics and exploratory plots indicate that, as in the SNOTEL case, the estimated series for the raw data provide valid substitution for the NOAA data sets. Exploratory analysis identified a yearly cycle in the data, which is captured by two of the three time domain models. The AR(3) weekly model fails to identify a significant yearly coefficient, but the AR(1)(1)24 and AR(1)(1)12 models for the semi–monthly and monthly series includes the yearly term. Frequency domain analysis identifies the 12–month cycle as the strongest contributor to observed variation, with 6– and 3–month cycles also playing a role.

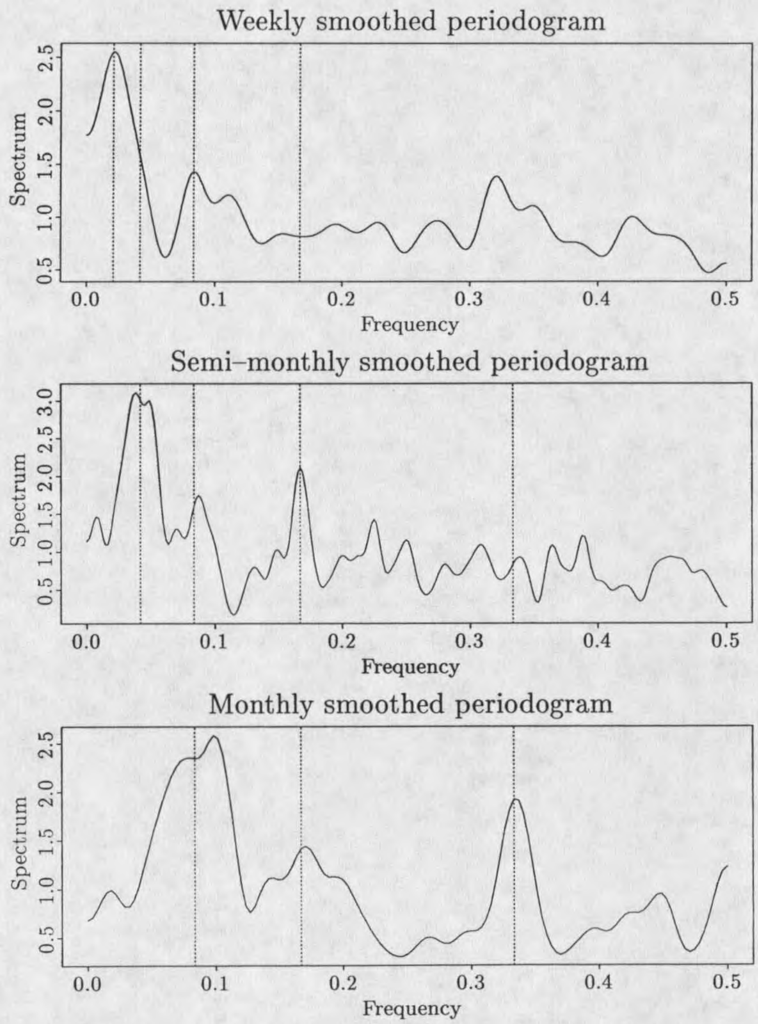

Figure 28: Smoothed periodograms of the NOAA estimated data sets. Dashed reference lines identify the frequencies with 12–, 6–, 3–, and 1.5–month periods.

CHAPTER 7

BIVARIATE DATA ANALYSIS

The univariate analyses of the data from each site result in different characteristics of the ARIMA models, but similarities in the spectral behavior at the two sites. Examining the data corresponding to the intersection of the time domains from each data set will complement the univariate analyses and provide additional understanding of the relationships between precipitation at the two locations.

The bivariate analysis begins by comparing the general behavior of the two series through the traces shown in Figure 29. A constant of 0.08 is added to the SNOTEL trace so that the two traces don't overlap. One can see the similarities in the general behavior of the two series without comparing individual data values. The two series share general patterns of highs and lows. The SNOTEL data set contains more individual extreme events not matched in the NOAA data set, but overall, they follow very similar patterns.

Another method of examining how the series behave together in time is to plot them as paired observations in a two dimensional scatterplot as shown in Figure 30. The line $y = x$ and the least squares line have been plotted as reference lines. The slopes of each of the least squares fits are 0.460, 0.488, and 0.620 for the weekly, biweekly, and monthly data. All slope coefficients are significant, and correlation coefficients are 0.473, 0.467, and 0.566. These results indicate a positive linear association between precipitation at the two sites and provide evidence that frequency domain analysis may uncover common component frequencies.

The bivariate frequency domain analysis begins by examining the smoothed periodogram overlays for each averaging period. These are displayed in Figure 31.

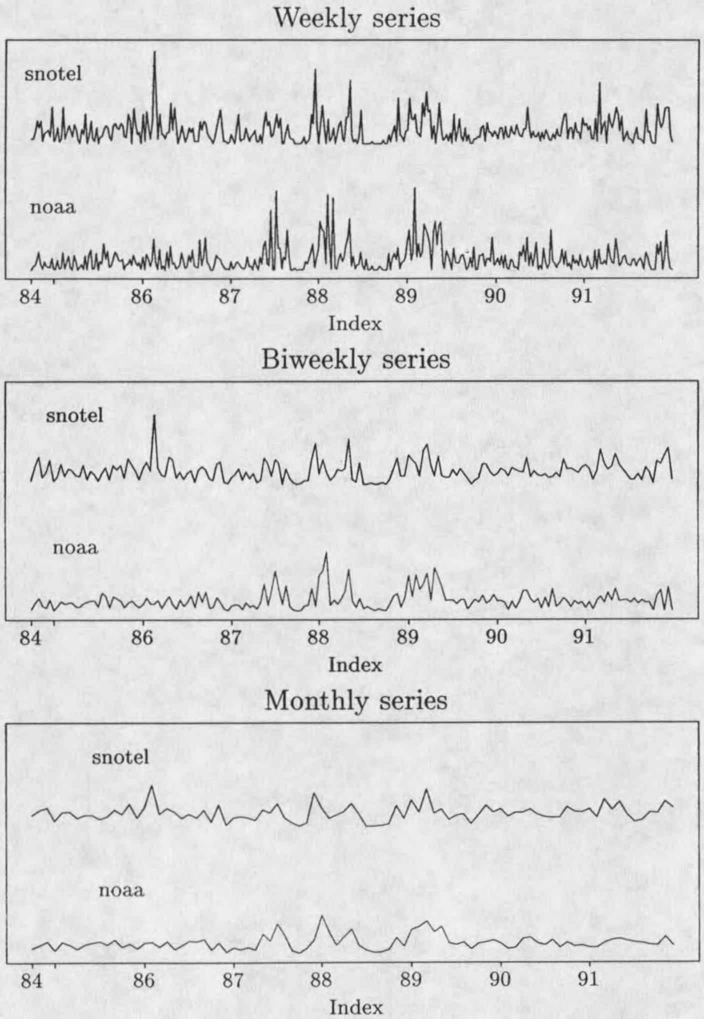

Figure 29: SNOTEL and NOAA weekly, biweekly, and monthly data sets plotted on the same time axis. Note that 0.8 has been added to the SNOTEL series to keep the traces separate.

In each plot in Figure 31, the SNOTEL periodogram is represented by a dashed line and the NOAA periodogram a solid one. Both spectral density estimates are calculated from the data corresponding to the intersection of the time domains. The periodograms illustrate very similar behavior patterns. Low frequencies dominate both data sets, as seen by the rather severe decrease in power with higher frequencies. Figure 31 indicates that the SNOTEL and NOAA data sets share component frequencies over all averaging periods. Because they show this relationship, the next step in the frequency domain analysis is to examine the shared component frequencies and compare their amplitudes and phases.

Figure 32 shows the coherence and phase spectral estimates for each set of averaged data. Coherence plots for each averaging period are shown in the first column, their corresponding phase estimates in the second. Coherence graphs show the squared correlation between amplitudes at each frequency. These coherence plots show high coherence at the component frequencies, and other frequencies as well. However, the coherence between the component frequencies is most important. The phase spectrum is interpretable only at frequencies with large coherence, and only the component frequencies are of interest.

The phase estimates are standardized and bounded by the interval $[0, 2\pi)$. Estimates near zero or 2π indicate no phase shift, while ordinates near π indicate a phase shift of half a period. Other ordinates have similar interpretations. In the phase plots of Figure 32 the estimates at those frequencies with large coherence, and most importantly, the component frequencies, are all near 2π, indicating no phase shift between the two series. In other words, there is no time lag between precipitation events at the two sites, at least on the scales measured. Observations on a smaller scale, such as daily, may show a lag, but these data sets indicate none.

The series traces and scatterplots show similar patterns in each series in the

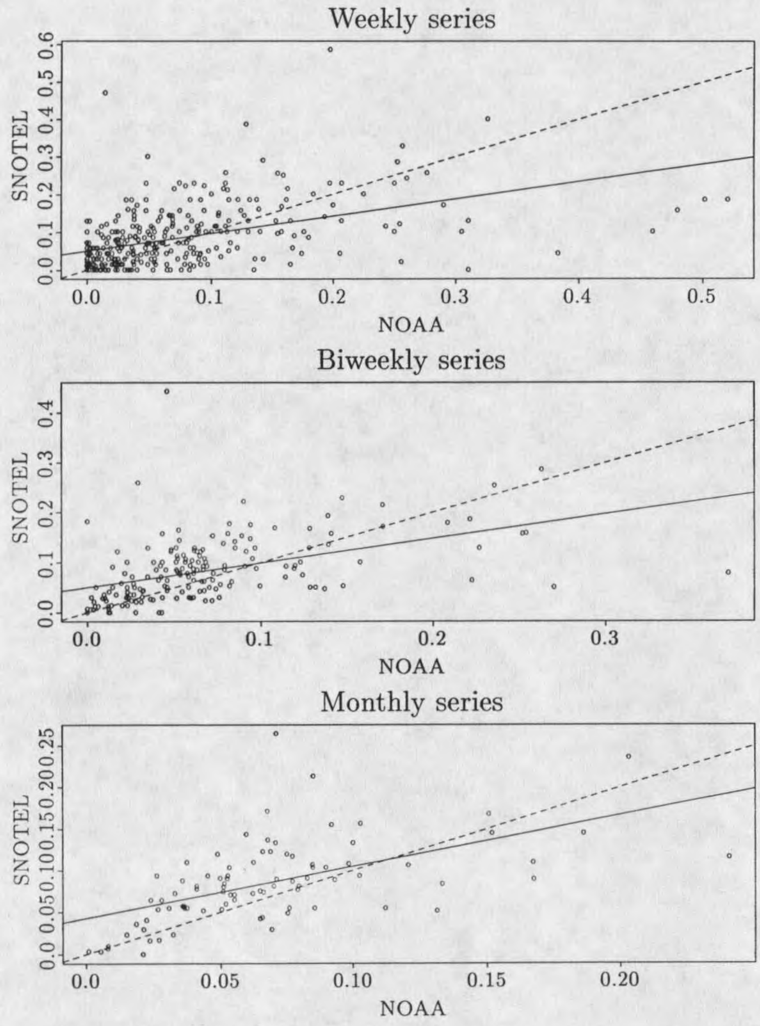

Figure 30: SNOTEL and NOAA data sets plotted against one another. The fitted least square regression appears as a solid line, the line $y = x$, dashed.

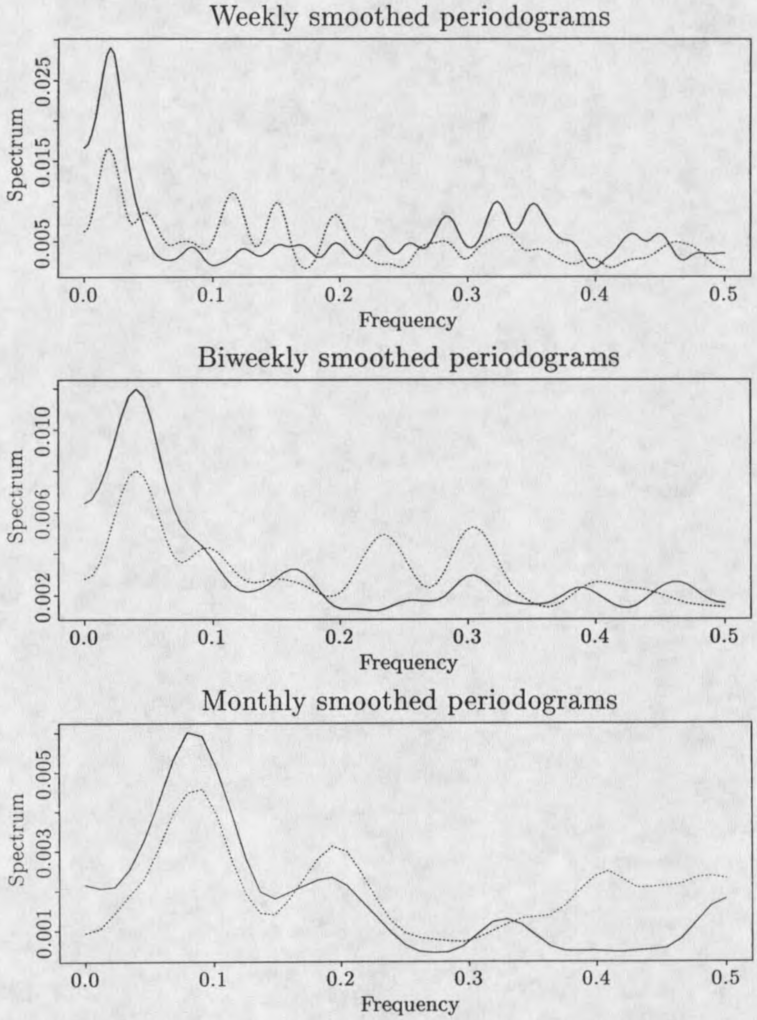

Figure 31: SNOTEL and NOAA spectral density estimates for all averaging periods. The dashed line represents the SNOTEL estimate, the solid line, the NOAA estimate.

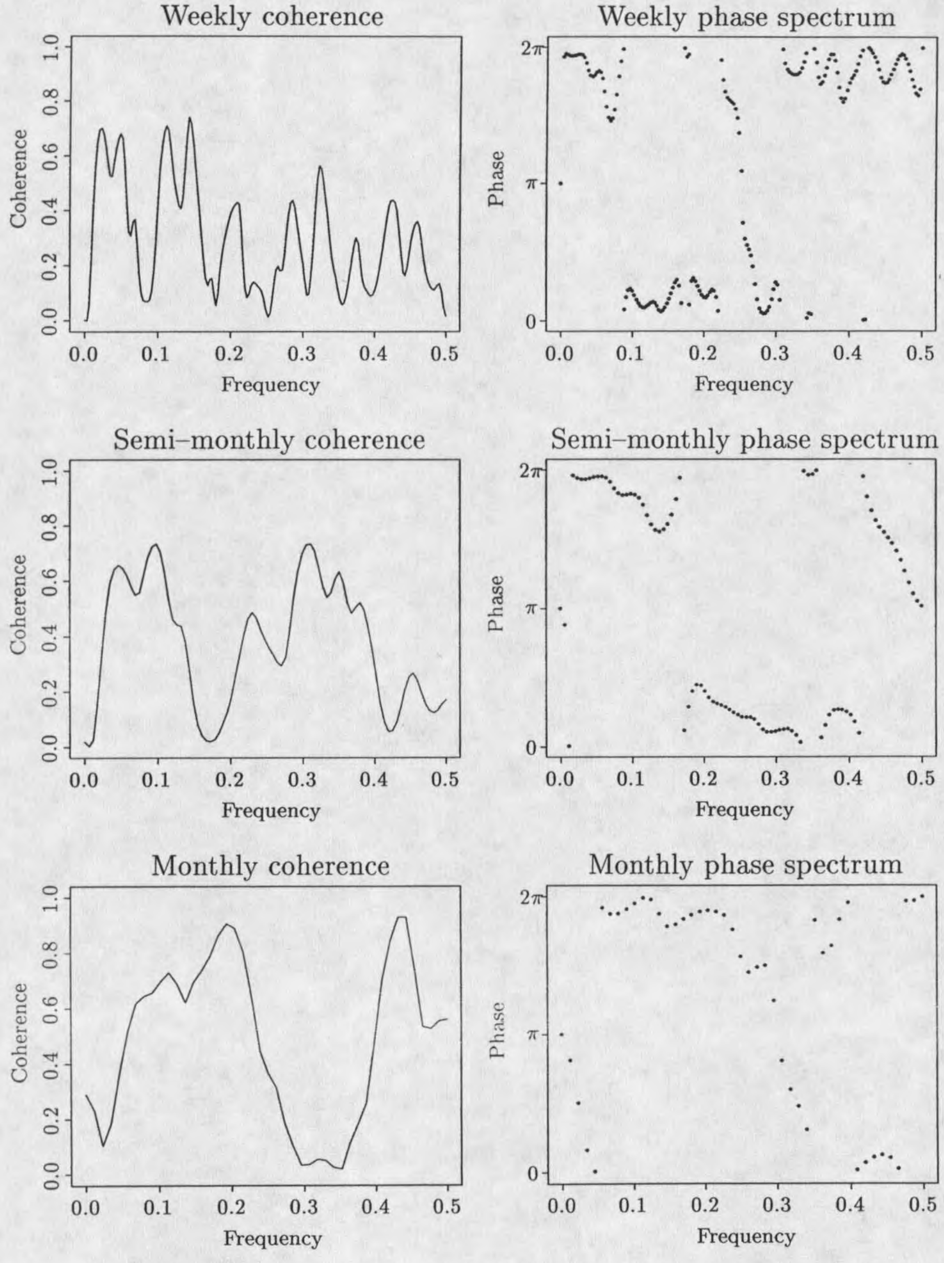

Figure 32: Coherence and phase estimates for the bivariate series. The first column gives coherence estimates for weekly, semi–monthly, and monthly series, the second column their corresponding phase estimates.

time domain, and the similarities carry over to the frequency domain. The underlying frequencies driving the time series at each location are components for both. They have highly correlated amplitudes and cycle at the same phases. The differences in location, altitude, and measuring equipment do not result in large differences in amounts or patterns of precipitation at the two sites.

CHAPTER 8

SUMMARY AND CONCLUSIONS

Exploratory analysis illustrates that when the missing observations are replaced by corresponding median values, the qualitative differences between the raw and estimated data sets are minimal. Both the SNOTEL and NOAA estimated data sets prove adequate for analysis and comparison.

Exploratory analysis also identified potential cyclic behavior in the data from both sites. The larger period–to–period variation in the SNOTEL data allowed easier identification of cyclic behavior, but time domain models seem to indicate the opposite.

Modeling the SNOTEL and NOAA data sets individually brings to light several interesting points of comparison. The NOAA data is successfully modeled as a yearly process while the SNOTEL models include no yearly component. The NOAA data gets better behaved in a modeling sense as the averaging period increases, while the SNOTEL data degenerates into a white noise sequence. Although both data sets contain a large noise component, in general, the NOAA data is less noisy than the SNOTEL data. Differences in location or surrounding terrain may account for the differing noise components in each of the data sets. The different measuring devices and the length of the record at each site may also contribute to differences in the time domain models.

When only the data from the intersection of the time domains is considered, the data sets show similar patterns both in the time and frequency domains. Although actual corresponding values may differ, in some cases extremely, overall patterns are similar. The two series exhibit quite similar frequency domain behavior–the yearly and semi–annual components are important to both. The lack of phase shift at either of the corresponding component frequencies indicates virtually no time lag between precipitation events at the two sites. Although predicting the amount of precipitation that falls in a given period at one site from the amount that falls at the other might prove very difficult, one could safely say that if it's precipitating at one site, it's also at the other, and vice versa.

Similarities in the periodic behavior between the two leads to the conclusion that when considering the SNOTEL data set for a tree–ring model, the NOAA set performs similarly. In fact, the NOAA set may be preferable because it contains a longer record. However, the question of whether either data set would provide an adequate basis for a tree–ring model is not determined, and is beyond the scope of this paper.

REFERENCES CITED

[1] P. Bloomfield. *Fourier analysis of time series.* John Wiley & Sons, 1976.

[2] G. E. P. Box and G. M. Jenkins. *Time series analysis: forecasting and control.* Holden–Day, 1976.

[3] G. E. P. Box and D. S. Pierce. Distribution of residual autocorrelations in arima time series models. *Journal of the American Statistical Association*, 65:1509–26, 1970.

[4] William S. Cleveland. *The elements of graphing data.* Wadsworth Advanced Books & Software, 1985.

[5] P. J. Diggle. *Time series: a biostatistical introduction.* Oxford University Press, 1990.

[6] H. C. Fritts. *Tree rings and climate.* Academic Press, 1976.

[7] J. M. Gottman. *Time series analysis: a comprehensive introduction for social scientists.* Cambridge University Press, 1981.

[8] L. H. Koopmans. *The spectral analysis of time series.* Academic Press, 1974.

[9] R. McCleary and R. A. Jr. Hay. *Applied time series analysis for the social sciences.* Sage, 1980.

[10] W. F. Quimby. Class notes: time series and spatial data analysis, September 1993.

[11] StatSci, a division of MathSoft, Inc. *Statistical Sciences, Inc. S-PLUS Reference Manual, Version 3.0*, 1992.

[12] John W. Tukey. *Exploratory data analysis.* Addison–Wesley Publishing Company, 1977.

APPENDIX

This appendix contains various statistics calculated for the SNOTEL and NOAA time domain models. The first column of each table identifies the series and model coefficients. The second column contains coefficient estimates and the third column contains only correlation estimates that show significant correlation between coefficients. The coefficients in question are specified in parentheses. The fourth and fifth columns correspond to significance tests for the last coefficient in the model. The last column gives values of the residual mean square(RMS) calculated from equation (2.15).

Table 5: Table of statistics for SNOTEL candidate models

Model	Coef	Corr	t–stat	p–val	RMS
Weekly series					
AR(1)					
$\hat{a}_1$	0.2209		4.25	0.000	18.762
AR(2)					
$\hat{a}_1$	0.2066				
$\hat{a}_2$	0.0603	-0.219	1.13	0.259	18.735
ARMA(1,1)					
$\hat{a}_1$	0.4970				
$\hat{b}_1$	0.2921	0.139	1.37	0.172	18.762
Semi–Monthly series					
Model	Coef	Corr	t–stat	p–val	RMS
AR(1)					
$\hat{a}_1$	0.1920		2.59	0.010	13.267
AR(1)(1)15					
$\hat{a}_1$	0.1948				
$\hat{a}_{15}$	-0.1847		-2.39	0.018	12.689
AR(2)					
$\hat{a}_1$	0.2030				
$\hat{a}_2$	-0.0394	-0.195	-0.552	0.602	13.229
ARMA(1,1)					
$\hat{a}_1$	-0.2410				
$\hat{b}_1$	-0.4562	0.198	-1.65	0.101	13.267
Monthly series					
Model	Coef	Corr	t–stat	p–val	RMS
AR(1)					
$\hat{a}_1$	0.1129		1.07	0.289	9.381
AR()(1)12					
$\hat{a}_{12}$	0.854		0.78	0.438	8.775
ARMA(1,1)					
$\hat{a}_1$	0.2203				
$\hat{b}_1$	0.1043	0.621	0.12	0.904	9.381

Table 6: Table of statistics for NOAA candidate models

Weekly series					
Model	Coef	Corr	t–stat	p–val	RMS
AR(3)					
$\hat{a}_1$	0.1351				
$\hat{a}_2$	0.0453	-0.139(1,2)			
$\hat{a}_3$	0.0999		3.33	0.001	33.181
AR(4)					
$\hat{a}_1$	0.1314				
$\hat{a}_2$	0.0435	-0.134(1,2)			
$\hat{a}_3$	0.0952				
$\hat{a}_4$	0.0344	-0.134(3,4)	1.14	0.254	33.166
AR(1)(1)48					
$\hat{a}_1$	0.1483				
$\hat{a}_{48}$	0.0440		1.43	0.153	32.481
ARMA(1,1)					
$\hat{a}_1$	0.7914				
$\hat{b}_1$	0.6719	0.054	0.824	0.410	33.211

Semi–monthly series					
Model	Coef	Corr	t–stat	p–val	RMS
AR(1)					
$\hat{a}_1$	0.1954		4.68	0.000	23.473
AR(1)(1)24					
$\hat{a}_1$	0.1925				
$\hat{a}_{24}$	0.0953		2.20	0.028	22.956
AR(2)					
$\hat{a}_1$	0.1742				
$\hat{a}_2$	0.0779	-0.189	1.83	0.068	23.452
ARMA(1,1)					
$\hat{a}_1$	0.5751				
$\hat{b}_1$	0.3978	0.110	2.62	0.009	23.473

Monthly series					
Model	Coef	Corr	t–stat	p–val	RMS
AR(1)					
$\hat{a}_1$	0.2013		3.41	0.001	16.583
AR(1)(1)12					
$\hat{a}_1$	0.2270				
$\hat{a}_{12}$	0.1861		3.07	0.002	16.217
ARMA(1,1)					
$\hat{a}_1$	0.1808				
$\hat{b}_1$	-0.0712	0.208	0.24	0.813	16.583