MODELING A FROST INDEX IN KANSAS,USA

by

YANG WANG

B.S., University of Science and Technology of China, 2004

A REPORT

submitted in partial fulfillment of the
requirements for the degree

MASTER

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2014

Approved by:

Major Professor
Perla Reyes

Copyright

Abstract

A frost index is a calculated value that can be used to describe the state and the changes in the weather conditions. Frost indices affect not only natural and managed ecosystems, but also a variety of human activities. In addition, they could indicate changes in extreme weather and climate events. Growing season length is one of the most important frost indices. In this report, growing season lengths were collected from 23 long-term stations over Kansas territory. The records extended to the late 1800s for a few stations, but many started observations in the early 1900s. Though the start dates of the records were different, the end dates were the same (2009).

To begin with, time series models of growing season length for all the stations were fitted. In addition, by using fitted time series models, predictions and validation checking were conducted. Then a regular linear regression model was fitted for the GSL data. It removed the temporal trend by doing regression on year and it showed us the relationship between GSL and elevation.

Finally, based on a penalized likelihood method with least angle regression (LARS) algorithm, spatial-temporal model selection and parameter estimation were performed simultaneously. Different neighborhood structures were used for model fitting. The spatial-temporal linear regression model obtained was used for interpreting growing season length of those stations across Kansas. These models could be used for agricultural management decision-making and updating recommendations for planting date in Kansas area.

Table of Contents

Table of Contents iv

List of Figures vi

List of Tables viii

Acknowledgements x

1 Introduction 1

1.1 Antecedents & Data Description 1

1.2 Methodology 2

2 Time Series Model 4

2.1 Model definition 4

2.2 Results 8

2.3 Conclusion 13

3 Linear regression model 14

3.1 Model Definition 15

3.2 Results 16

3.3 Conclusion 17

4 Spatial Temporal Model 19

4.1 Model Definition 19

4.2 Defining the Neighborhood sturcture 21

4.3 Model fitting 23
4.4 Results and conclusion 25

5 Discussion **31**

Bibliography **33**

A Time series analysis 18 remaining stations **35**

B Some details of penalized likelihood method **40**

List of Figures

1.1 Figure1.1 Elevation map of 23 long-term weather stations in Kansas 3

2.1 Figure 2.1 Time Series plots 9

2.2 Figure 2.2 Forcasting and validation for GSL at Saint Francis 11

2.3 Figure 2.3 Forcasting and validation for GSL at Elkhart 11

2.4 Figure 2.4 Forcasting and validation for GSL at Horton 12

2.5 Figure 2.5 Forcasting and validation for GSL at Columbus (continuous line prediction. Dashed line 95% C.I.. Green line true values for validation) . . . 12

2.6 Figure 2.6 Forcasting and validation for GSL at Larned No.2 (continuous line prediction. Dashed line 95% C.I.. Green line true values for validation) . . . 13

3.1 Figure 3.1 GSL vs elevation, latitude, longitude 15

3.2 Figure 3.2 ACF and PACF of regression residuals Atchison station 17

3.3 Figure 3.3 Bubble plot of linear model residuals 18

4.1 Figure 4.1 Bubble plot of GSL 20

4.2 Figure 4.2 Neighborhood partition plot 24

A.1 Figure A.1 Time series plots. Atchison and Ashland 35

A.2 Figure A.2 Time series plots. Colby 1sw and Ft Scott 36

A.3 Figure A.3 Time series plots. Hays 1s and Independence 36

A.4 Figure A.4 Time series plots. Lakin and Manhattan 37

A.5 Figure A.5 Time series plots.Mcpherson and Medicine Lodge 37

A.6 Figure A.6 Time series plots. Minniapolis and Winfield 3Ne 38
A.7 Figure A.7 Time series plots. Oberlin and Ottawa 38
A.8 Figure A.8 Time series plots. Phillipsburg and Sedan 39
A.9 Figure A.8 Time series plots. Tribune 1w and Wakeeney 39

List of Tables

2.1 Time Series models of GSL for 23 Kansas centennial stations (1908-2009) . . 10

3.1 Estimates of regular linear regression model 16

4.1 Fitted spatial temporal model for centered GSL, 5 time lags and different neighborhood structures: $CaseI$, 120 miles, 200 miles and 300 miles radius partition of the neighborhoods. $CaseII$ same first 3 order neighborhoods as $CaseI$ and the combined neighbors inside the 120&200 mile circle as the fourth neighborhood. $CaseIII$ 120 miles, 200 miles, 300 miles, 400 miles radius partition of the neighborhood. $CaseIV$ 150 miles, 250 miles, 350 miles, 450 miles radius. $CaseV$ combined $CaseIII$ and $CaseIV$, and added neighbors within 200 miles radius. 26

4.2 Fitted spatial temporal model for centered GSL, 10 time lags and different neighborhood structures: $CaseI$, 120 miles, 200 miles and 300 miles radius partition of the neighborhoods. $CaseII$ same first 3 order neighborhoods as $CaseI$ and the combined neighbors inside the 120&200 mile circle as the fourth neighborhood. $CaseIII$ 120 miles, 200 miles, 300 miles, 400 miles radius partition of the neighborhood. $CaseIV$ 150 miles, 250 miles, 350 miles, 450 miles radius. $CaseV$ combined $CaseIII$ and $CaseIV$, and added neighbors within 200 miles radius. 27

4.3 Fitted spatial temporal model for standardized GSL, 5 time lags and different neighborhood structures: $CaseI$, 120 miles, 200 miles and 300 miles radius partition of the neighborhoods. $CaseII$ same first 3 order neighborhoods as $CaseI$ and the combined neighbors inside the 120&200 mile circle as the fourth neighborhood. $CaseIII$ 120 miles, 200 miles, 300 miles, 400 miles radius partition of the neighborhood. $CaseIV$ 150 miles, 250 miles, 350 miles, 450 miles radius. $CaseV$ combined $CaseIII$ and $CaseIV$, and added neighbors within 200 miles radius. 28

4.4 Fitted spatial temporal model for standardized GSL, 10 time lags and different neighborhood structures: $CaseI$, 120 miles, 200 miles and 300 miles radius partition of the neighborhoods. $CaseII$ same first 3 order neighborhoods as $CaseI$ and the combined neighbors inside the 120&200 mile circle as the fourth neighborhood. $CaseIII$ 120 miles, 200 miles, 300 miles, 400 miles radius partition of the neighborhood. $CaseIV$ 150 miles, 250 miles, 350 miles, 450 miles radius. $CaseV$ combined $CaseIII$ and $CaseIV$, and added neighbors within 200 miles radius. 29

Acknowledgments

I would like to express sincere thanks to my major professor Dr. Perla Reyes for her help and advice through the report process and all my master study in the program. It would not have been possible to complete this work without her guidance and encouragement. I learned a lot from her. I would also like to give many thanks to my report committee members, Dr. Christopher Vahl and Dr. Weixin Yao for their time and help. Specifically, I would like to thank Aavudai Anandhi for providing the data and her suggestions for the research. I am very grateful to all the faculties of Statistics department in Kansas State University as well.

My family members and friends have substantively contributed to this work. I would like to thank my wife for her support during my research and study. And I would like to thank my parents for their encouragement during my graduate education. Finally, I would like to thank my lovely daughter for making me happy every day.

Chapter 1

Introduction

Growing season length is one of the most important frost indices. It is used to plan agricultural activities. For instance, a longer growing season could allow farmers to diversify crops or have multiple harvests from the same plot. In this chapter, the data and techniques that will be used for analyzing it will be introduced.

1.1 Antecedents & Data Description

In the United States, agriculture is a very important industry. It is not only an important food source for people all over the U.S., but also a large part of the national economy (USGCRP 2009). According to the U.S. census of agriculture in 2012, 2.1 million farms are distributed across the country with total area of 915 million acres (U.S. Census 2012). Particularly, most of the agriculture activities are concentrated in the central part of America (Hatfield, J. 2012).

Kansas state, which is at the right center of the U.S., is known as the "Wheat State" and "Breadbasket of the World" (Kansaspedia 2012). In Kansas, farming has become a way of life. Agriculture impacts in politics, culture, social customs, laws, as well as traditions. Kansas economy relies significantly on agriculture related businesses.

It is well known that agriculture is highly dependent on climate conditions. When climate changes, practices and technologies should be changed also. Frost indices, which are commonly used in agricultural industry, can be indicative of changes in extreme climate events (Meehl 2004). Some extreme conditions can increase the risk of natural disasters and they affect nature and humans in many aspects. One of those important frost indices is Growing season length (GSL). It is useful in determining crop cycle lengths and calendars. In this report, it was calculated as the difference between last spring frost (LSF) and first fall frost (FFF). Here, LSF is defined as the day in March through May with minimum air temperature (T_{min}) $< 0\,^{\circ}$C for the last time until fall. And FFF is defined as the day in September through November with $T_{min} < 0\,^{\circ}$C for the first time since spring. In Anandhi et.al (Anandhi 2013), increasing linear trends were observed for GSL at 23 centennial weather stations over Kansas.

Information from 23 centennial weather stations in Kansas state are distributed across Kansas (Figure 1.1). Anandhi et.al[5] downloaded daily minimum air temperature data of these 23 stations from the High Plains Regional Climate Center (HPRCC)s website and calculated LSF, FFF as well as GSL base on the definition provided above. In this report, GSL data for these stations will be analyzed. The data extended to the late 1800s for a few stations, but the majority started in the early 1900s. The end year was the same for all stations, 2009. Figure 1.1 shows the distribution of the stations across Kansas.

In addition, latitude, longitude and elevation of stations were used to model GSL. The result of an ordinary linear regression model is presented in Chapter 3 and a spatial-temporal linear model is described in Chapter 4.

1.2 Methodology

Growing season is the period of the year during which growing conditions for indigenous vegetation and cultivated crops are most favourable. Initially, the GSL times series of each

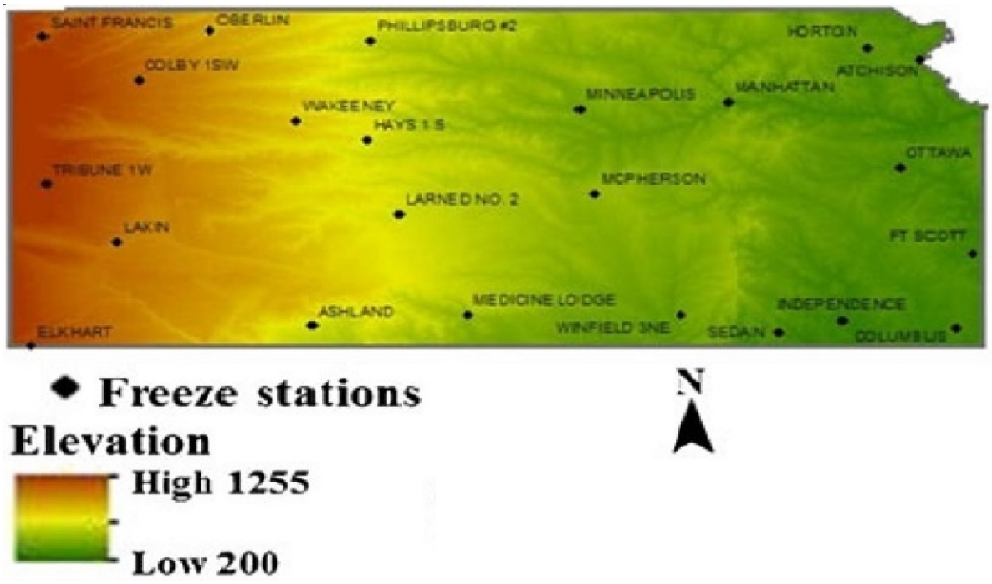

Figure 1.1: *Elevation map of 23 long-term weather stations in Kansas, US*

station was analyzed independently. Autoregressive integrated moving average (ARIMA) models and Seasonal autoregressive integrated moving average (SARIMA) models were fitted for each stations and prediction was conducted in Chapter 2. Additionally, each station's model was validated. The last 6 years points were saved and compared with the predicted values of those 6 year.

In Chapter 3, for simplicity, the data was balanced by constricting GSL for all stations to start in 1908. Considering time, latitude, longitude and elevation as covariates, then stepwise model selection with Akaike's information criterion (AIC) was used to determine the best linear regression model.

In Chapter 4, an algorithm devised for simultaneous spatial-temporal model selection and parameter estimation was applied to analyze GSL of all 23 stations in Kansas. Neighborhood structure was identified and a spatial-temporal linear model was fitted. These models considered the spatial temporal pattern and could be very useful for agricultural management decision-making and updating recommendations for planting date in Kansas area.

Chapter 2

Time Series Model

Time series is a an ordered sequence of values of a variable at equally spaced time intervals. It is widely used in areas such as econometrics, mathematical finance, weather forecasting and so on. In this chapter, definition of time series models will be firstly introduced. Then models are fitted and analyzed for each station. The conclusions of these time series models for stations is presented in the last section of this chapter.

2.1 Model definition

For this analysis, five different time series models were considered (Box 2013), autoregressive (AR), moving average (MA), autoregressive moving average (ARMA), autoregressive integrated moving average (ARIMA) and seasonal autoregressive integrated moving average (SARIMA). AR, MA and ARMA models are used for time series with lagged linear relations. ARMA model is a combination of AR and MA models. Adding nonstationarity to ARMA models leads to the ARIMA model. For time series with seasonal phenomenon, SARIMA models are commonly used. The definitions of these models will be introduced below. We will start with the backshift operator and the concept of stationary that would be used later.

Definition 1 Backshift operator is defined by

$$Bx_t = x_{t-1}$$

and extend it to powers $B^2x_t = B(Bx_t) = Bx_{t-1} = x_{t-2}$, and so on. Thus,

$$B^k x_t = x_{t-k}$$

Definition 2 For a process $\{X_t\}$, let $F_X(x_{t_1+\tau}, \ldots, x_{t_k+\tau})$ represent the cumulative distribution function of the joint distribution X_t at times $t_1 + \tau, \ldots, t_k + \tau$. Then $\{X_t\}$ is said to be strict stationary if, for all k, for all τ, and for all $t_1, \ldots, t_k$,

$$F_X(x_{t_1+\tau}, \ldots, x_{t_k+\tau}) = F_X(x_{t_1}, \ldots, x_{t_k}).$$

The process is said to be weakly stationary if $E(X_t)$ and $Var(X_t)$ does not depend on t, and $Cov(X_t, X_{t+k})$ only depends on lag k.

Definition 3 An autoregressive model of order p, abbreviated AR(p), is of the form

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots + \phi_p x_{t-p} + \omega_t$$

where x_t is stationary (joint probability distribution does not change when shifted in time), and ϕ_1, ϕ_2,..., ϕ_p are constants ($\phi_p \neq 0$). ω_t is a Gaussian white noise series (a stationary process having a normal distribution with mean zero and constant variance) with variance ${\sigma_\omega}^2$. The mean of x_t here is zero. If the mean, μ, of x_t is not zero, replace x_t by $x_t - \mu$,

$$x_t = \alpha + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots + \phi_p x_{t-p} + \omega_t$$

where $\alpha = \mu(1 - \phi_1 - \ldots - \phi_p)$. By using the backshift operator, this can be written as

$$(1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p) x_t = \omega_t,$$

or

$$\phi(B) = \omega_t$$

where $\phi(B)$ is the autoregressive operator which is defined as

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p.$$

Definition 4 The moving average model of order q, or MA(q) model, is defined to be

$$x_t = \omega_t + \theta_1 \omega_{t-1} + \theta_2 \omega_{t-2} + \ldots + \theta_q \omega_{t-q},$$

or

$$x_t = \theta(B)\omega_t,$$

where $\theta(B)$ is the moving average operator which is defined as

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \ldots + \theta_p B^p.$$

Definition 5 A time series $x_t, t = 0, \pm 1, \pm 2, \ldots$ is $ARMA(p, q)$ if it is stationary and

$$x_t = \phi_1 x_{t-1} + \ldots + \phi_p x_{t-p} + \omega_t + \theta_1 \omega_{t-1} + \ldots + \theta_q \omega_{t-q},$$

where $\phi_p \neq 0, \theta_q \neq 0$, and ${\sigma_\omega}^2 > 0$. The parameters p and q are called the autoregressive and the moving average orders, respectively. If x_t has a nonzero mean μ, set $\alpha = \mu(1 - \phi_1 -$

$\cdots - \phi_p$) and write the model as

$$\phi(B)x_t = \alpha + \theta(B)\omega_t.$$

In many situations, time series can be thought of as being composed of two components, a nonstationary trend component and a zero-mean stationary component. However, differencing such a process will lead to a stationary process. The integrated ARMA, or ARIMA, model is a broadening of the class of ARMA models to include differencing.

Definition 6 A process x_t is said to be $ARIMA(p, q, d)$ if

$$\phi(B)(1 - B)^d x_t = \theta(B)\omega_t,$$

where $\phi(B)$ and $\theta(B)$ are autoregressive operator and moving average operator. If $E((1 - B)^d x_t) = \mu$, the model can be written as

$$\phi(B)(1 - B)^d x_t = \delta + \theta(B)\omega_t,$$

where $\delta = \mu(1 - \phi_1 - \cdots - \phi_p)$.

When there are seasonal phenomena, the seasonal ARIMA model can be used.

Definition 7 The multiplicative seasonal autoregressive integrated moving average model, or SARIMA model is given by

$$\Phi_P(B^s)\phi(B)(1 - B^s)^D(1 - B)^d x_t = \delta + \Theta_Q(B^s)\theta(B)\omega_t,$$

where ω_t is the usual Gaussian white noise process and s is seasonal period. This model is generally denoted as $ARIMA(p, d, q) \times (P, D, Q)_s$. $\phi(B)$ and $\theta(B)$ are ordinary autoregressive and moving average operators of orders p and q. $\Phi_P(B^s)$ and $\Theta_Q(B^s)$, seasonal autoregressive and moving average operators of orders P and Q seasonal period s, are defined

as

$$\Phi_P(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps}$$

and

$$\Theta_Q(B^s) = 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \cdots + \Theta_Q B^{Qs}.$$

2.2 Results

Model fitting for all stations are similar. For illustration, selected stations Saint Francis, Columbus, Elkhart, Larned No.2, and Horton, are shown in Figure 2.1.

After having all time series plots of GSL, ACF, PACF plots and plots of ACF and PACF with different seasonal periods of these time series were used to select candidate models. Among candidate models for each station, the one with smallest AIC has been selected as the final model. Fitted models of each stations are listed in Table 2.1.

From Table 2.1, it can be seen that GSL of Saint Francis station and Larned NO.2 station are modeled by seasonal ARIMA model. GSL of Station ELkhart is modeled by ARIMA model with high orders. GSL of Horton station is modeled by ARIMA model with low orders and GSL of Columbus station is almost a white noise. The results indicate that GSL of stations with higher elevation tend to have higher lagged correlation while GSL of stations with lower elevation tend to have lower lagged correlation.

After having models of each station, prediction of GSL can be conducted. Time series plots, forcasting and validation plots for the other stations are listed in Appendix A. Figures 2.2-2.6 show predictions of the 5 selected stations as well as the validation of results.

In forcasting plots (Figure 2.2), data from all years are used and predictions for next 6 years' GSL are based on the models. Solid lines are the predicted values of GSL for next 6 years and two dashed lines mark the limits of 95% confidence intervals for the predicted values. In validation plots from Figure 2.2, the last 6 years data are set aside for checking the rest of the data are used for model fitting. The red solid lines are the predicted values

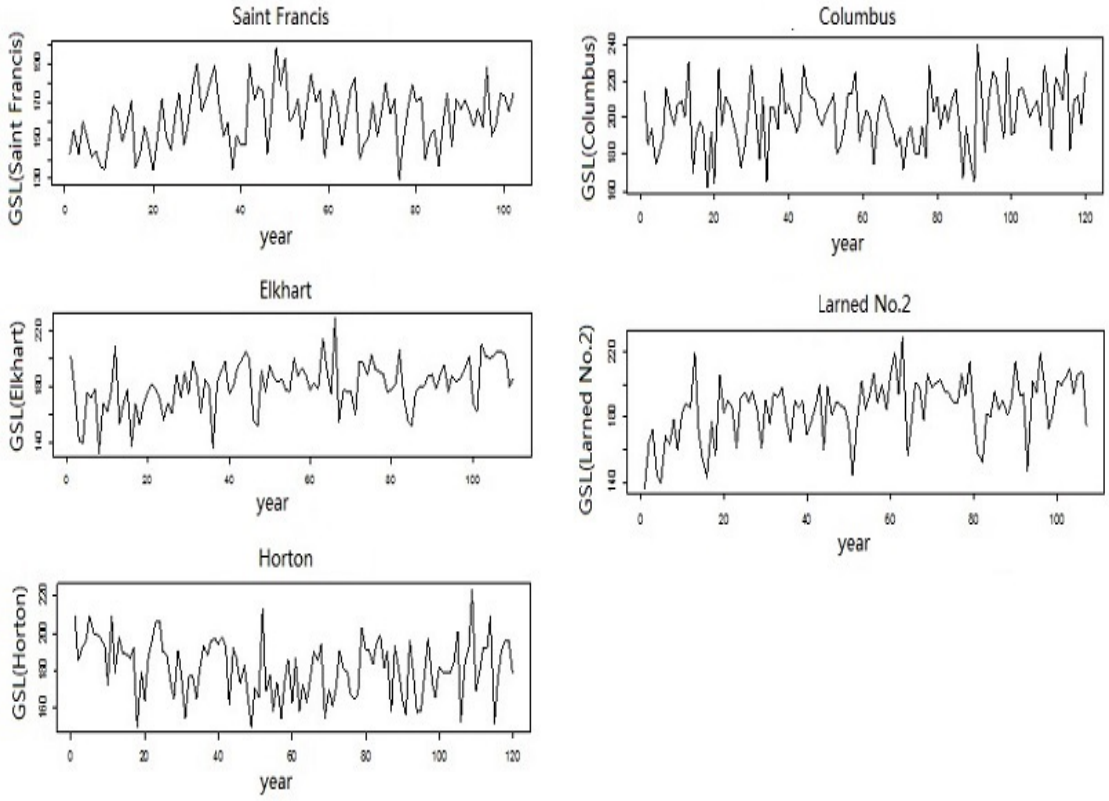

Figure 2.1: *Time Series plots of 5 selected stations (Saint Francis, Columbus, Elkhart, Larned No.2, Horton)*

Station	COOPID	Time Series model of GSL
Saint Francis	147093	SARIMA (1,0,0) *(1,1,2) with S=8
Elkhart	147093	ARMA (10,10)
Horton	143810	ARMA(1,1)
Columbus	141740	White noise
Larned No2	140365	SARIMA (0,0,1)*(2,1,1) with S=11
Atchison	140405	ARMA(3,3)
Ashland	140365	ARMA(2,3)
Colby 1sw	141699	SARIMA (1,0,0)*(2,1,1) with S=15
Ft Scott	142835	ARIMA(1,1,1)
Hays 1S	143527	White noise
Independence	143954	ARIMA(5,1,4)
Lakin	144464	SARIMA (0,1,1)*(1,1,1) with S=19
Manhattan	144972	ARMA(1,1)
Mcpherson	145152	ARIMA(0,1,1)
Medicine lodge	145175	ARMA(1,1)
Minneapolis	145363	ARIMA(4,1,3)
Winfield 3Ne	148964	ARMA(3,3)
Oberlin	145906	SARIMA (2,0,0)*(2,1,1) with S=22
Ottawa	146128	MA(1)
Phillipsburg No2	146378	ARIMA(2,1,3)
Sedan	147305	ARIMA(0,1,1)
Tribune 1W	148235	SARIMA (1,0,0)*(1,1,1) with S=8
Wakeeney	148495	ARIMA(3,1,1)

Table 2.1: *Time Series models of GSL for 23 Kansas centennial stations (1908-2009)*

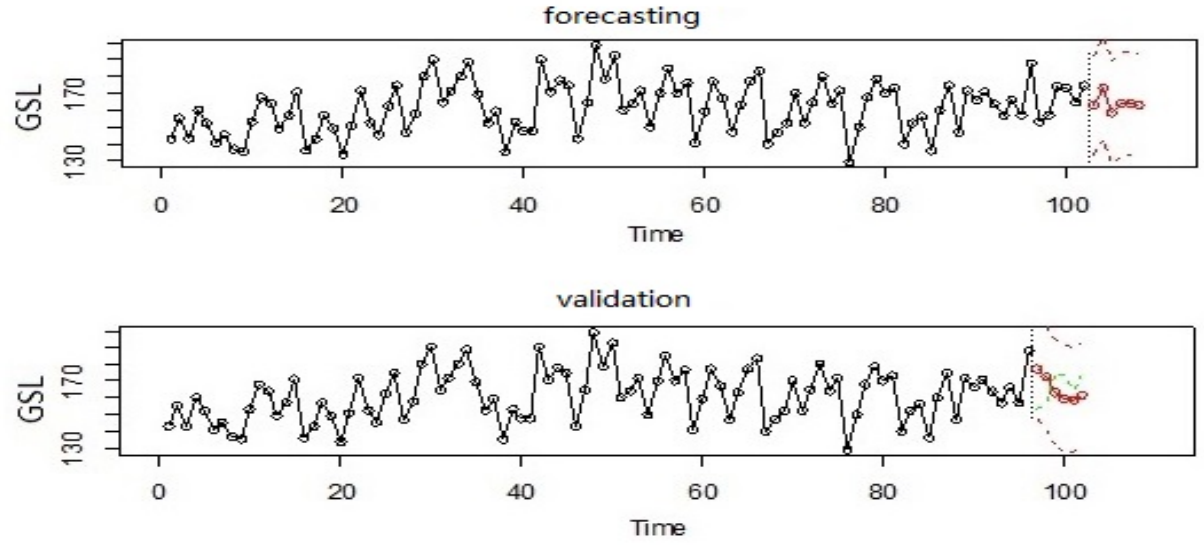

Figure 2.2: *Forcasting and validation Saint Francis station (continuous line prediction. Dashed line 95% C.I.. Green line true values for validation)*

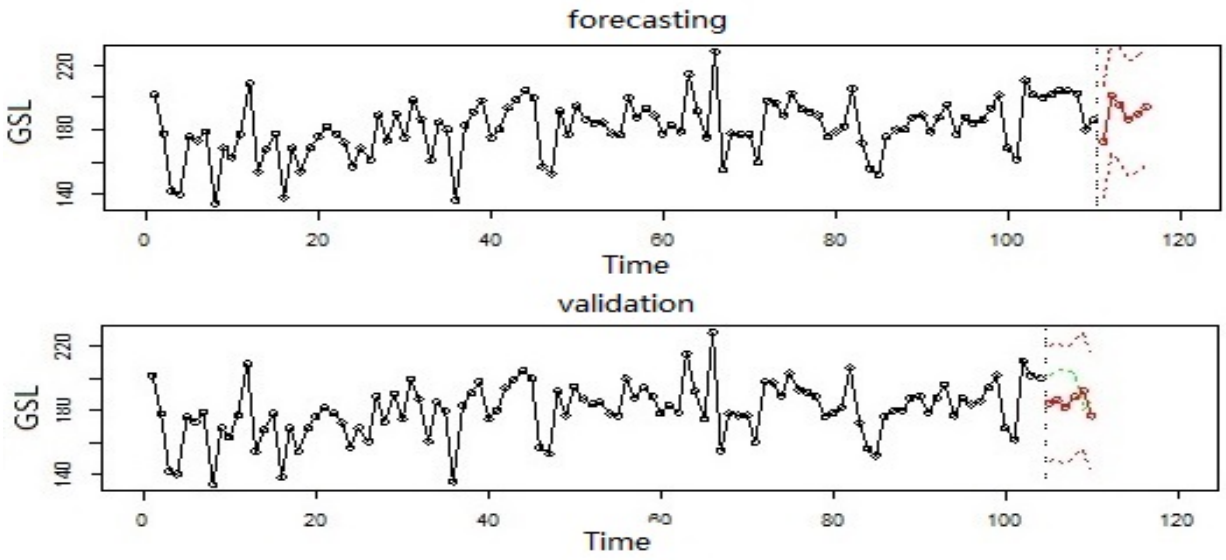

Figure 2.3: *Forcasting and validation Elkhart station (continuous line prediction. Dashed line 95% C.I.. Green line true values for validation)*

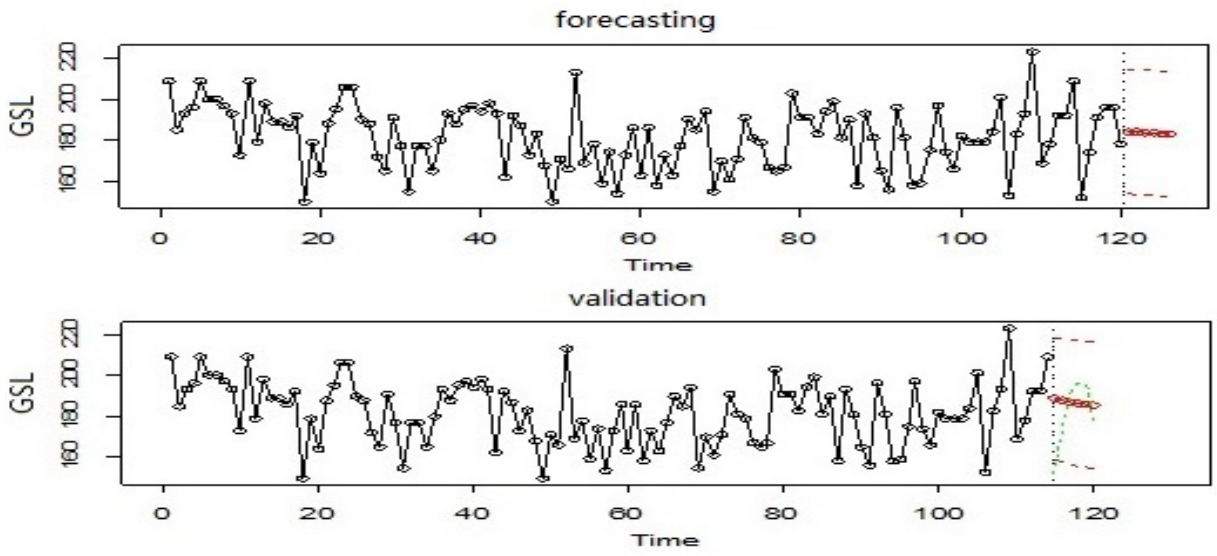

Figure 2.4: *Forcasting and validation Horton station (continuous line prediction. Dashed line 95% C.I.. Green line true values for validation)*

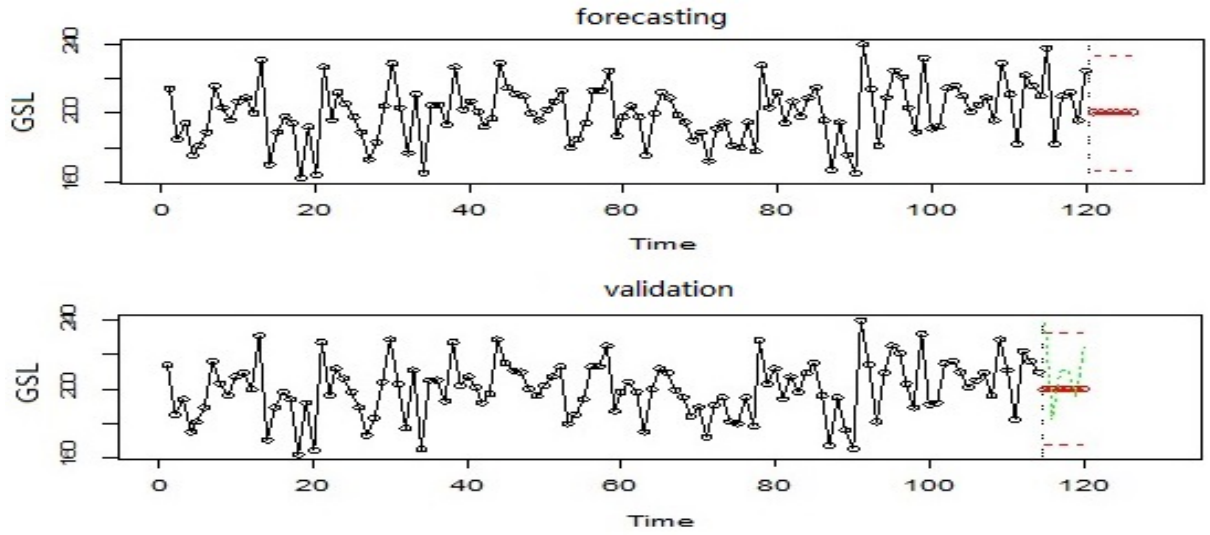

Figure 2.5: *Forcasting and validation Columbus station*

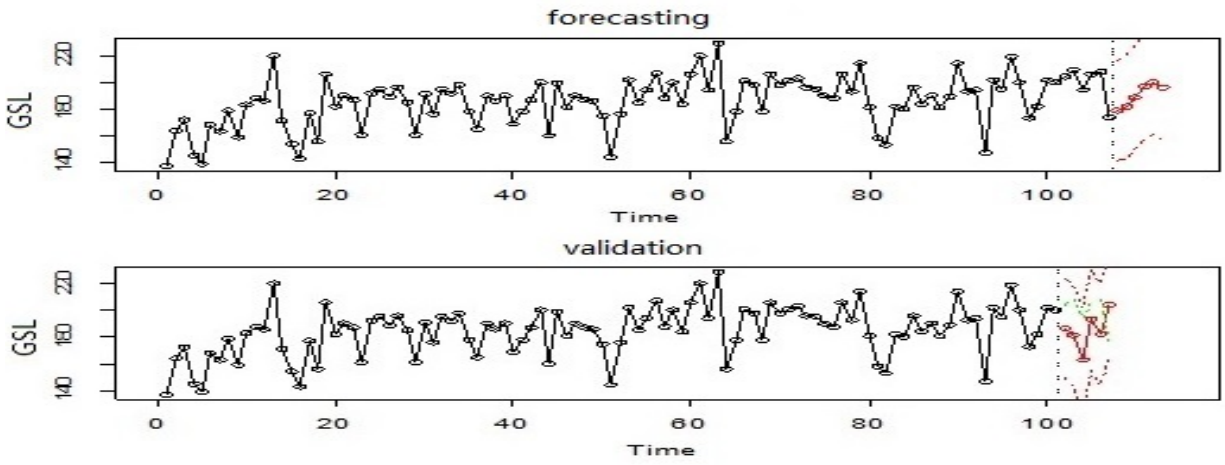

Figure 2.6: *Forcasting and validation Larned No.2 station*

of the last 6 years and the red dashed lines are the limits of its confidence intervals. The green dashed line marked the true values of GSL for the last 6 years.

2.3 Conclusion

From results obtained in last section, it can be seen that stations with higher elevations tend to have more complicated time series models. While stations with lower elevations tend to have relatively simple models. Thus, it is possible that elevation of station affects GSL. Another possibility is the presence of a spatial trend. Therefore, in the next chapter, linear regression on elevation, latitude and longitude will be conducted by using GSL data of all 23 long-term stations in Kansas.

The model checking indicate that all real values are close to the predicted values, or inside of the confidence intervals. Confidence intervals tend to be wide due to the high variability of the data.

Chapter 3

Linear regression model

According to the results in Chapter 2, it is possible that the growing season length is associated with the elevation of station or its location. Plots about the relationship between GSL and elevation, latitude, longitude and year are shown as Figure 3.1. It can be seen that there are obvious correlations. Actually, the correlation between GSL and elevation is -0.503, the correlation between GSL latitude is -0.412, the correlation between GSL and longitude is 0.505.

In this chapter, we present a linear regression model that used year, elevation, latitude and longitude as covariates. Because the length of GSL data for stations are not the same, data was truncated to make the length of GSL data for every station the same. The latest record year among 23 long-term stations is 1908, so the GSL analyzed in this chapter are all from 1908 to 2009. This truncation is not necessary for a linear model. However, the model introduced in Chapter 4 requires balanced data. We truncated the data to make both results comparable.

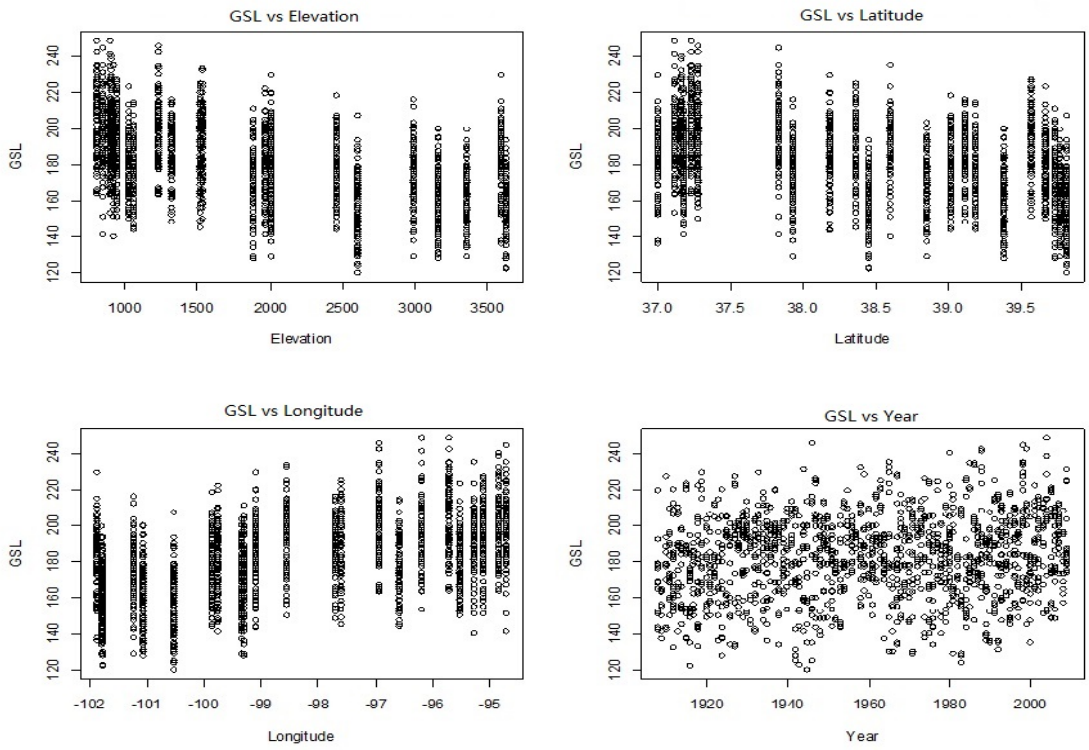

Figure 3.1: *GSL VS elevation, latitude, longitude*

3.1 Model Definition

Consider GSL as response and denote it as Y, then the linear regression model of GSL can be written as

$$Y_{it} = \beta_0 + \beta_1 t + \beta_2 X_{ielev} + \beta_3 X_{ilati} + \beta_4 X_{ilong} + interactions + \varepsilon_i t, \tag{3.1}$$

where Y_{it} denotes the GSL of ith station at year t. Four predictor variables here are year, elevation, latitude and longitude. t, X_{ielev}, X_{ilati}, X_{ilong} are the values of these four predictor variables for the ith station. To save space, the interaction term includes all the 2 way, 3 way and 4 way interactions of the four predictor variables. The parameters of the model are β's, σ^2, the error term is ε_{it}.

Coefficients	Estimate	Std. Error	P-value
Intercept	-1093	530	0.039
Year	0.098	0.012	<0.001
Elevation	-0.004	0.001	<0.001
Latitude	34.58	13.69	0.012
Longitude	-14.01	5.386	0.009
Latitude:Longitude	0.427	0.139	0.002
σ^2	276	-	-

Table 3.1: *Estimates of regular linear regression model*

3.2 Results

The results of the stepwise algorithm from R using AIC for model comparison are presented in Table 3.1.

It can be observed that elevation and year are significant. A significant interaction between latitude and longitude indicates the presence of a NW-SE trend. The coefficient of year is positive which indicates that GSL gets longer with year. GSL tends to decrease when elevation gets higher. The final regression model can be written as

$$GSL = -1093 + 0.098year - 0.004elev + 34.58lati - 14.01long + 0.427lati : long + \epsilon,$$

Having the regression model of GSL, the examination of the residuals for all stations showed no temporal trend. See for instance, the ACF and PACF plots of station at ATCHISON shown in Figure 3.2 where no obvious spikes can be found. Figure 3.3 shows us the bubble plots of the residuals of 6 randomly selected years. The sizes of the bubbles are proportional to the absolute value of the residuals. It is evident that the spatial trend is not explained by the linear model.

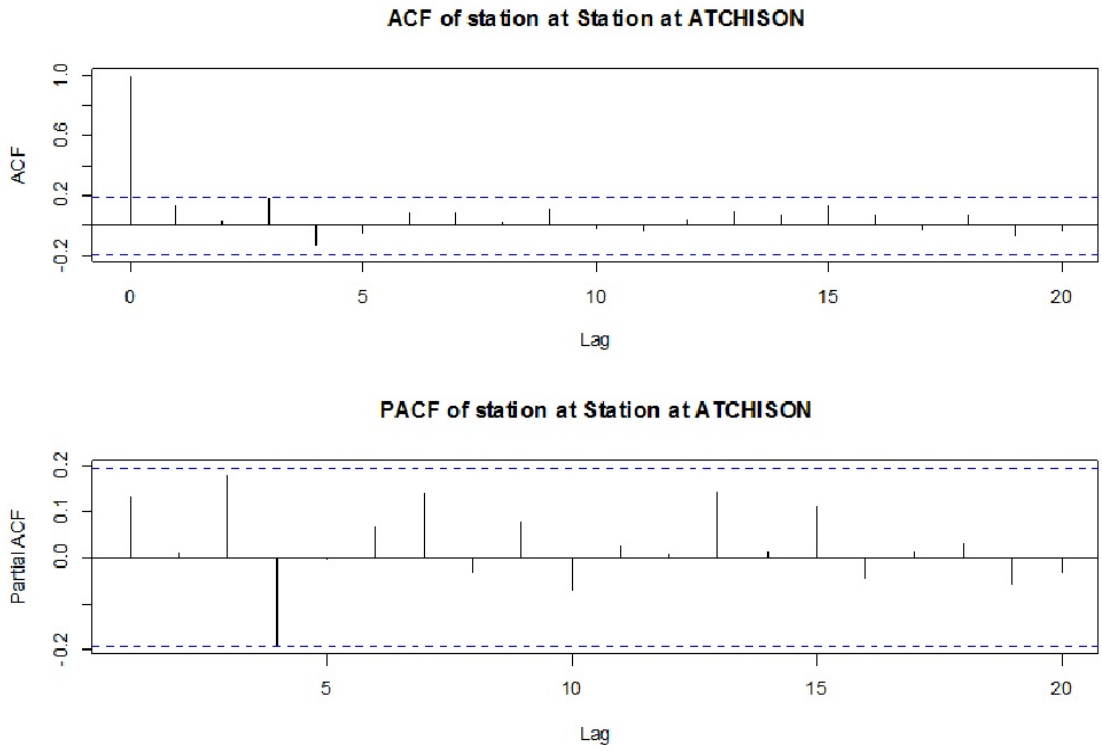

Figure 3.2: *ACF and PACF of regression residuals ATCHISON station*

3.3 Conclusion

By fitting a regular linear regression model, it can be seen that elevation and location of the station significantly affect GSL. For the effect of elevation, stations with higher elevation tend to have shorter GSL, while stations with lower elevation tend to have longer GSL. For the location of the station, stations in the southeast part of Kansas tend to have longer GSL and stations in the northwest part of Kansas tend to have shorter elevation.

From the examination of the residuals, it seems that including year may account for temporal trend. However, the spatial association is not completely explained. In the next chapter, a spatial-temporal linear regression model will be fitted by using the GSL data and the results will be interpreted and compared with the results in this chapter.

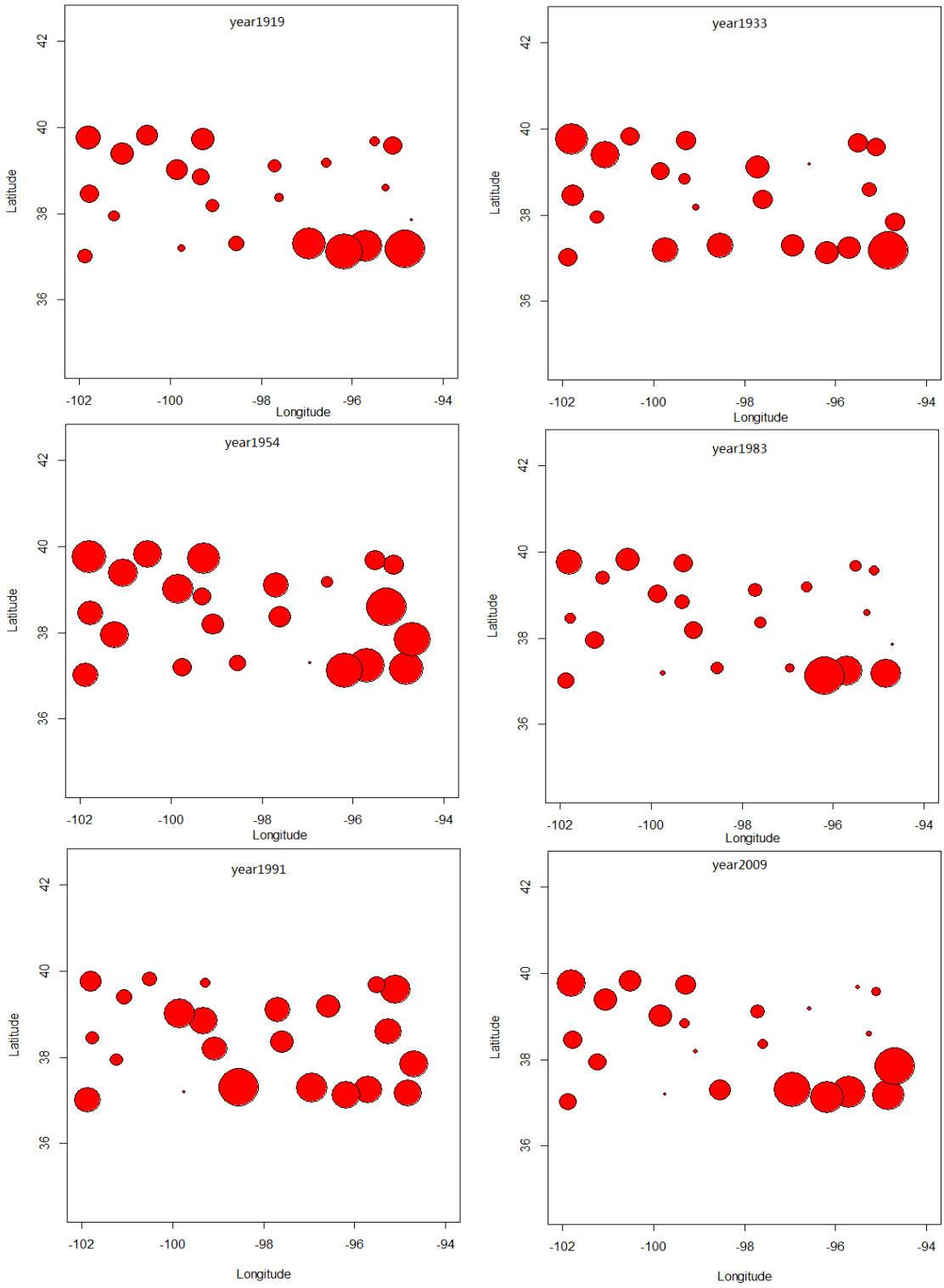

Figure 3.3: *Bubble plot of residuals in 6 randomly selected years. The size of the bubble is proportional to the residual's absolute value*

Chapter 4

Spatial Temporal Model

The results from Chapter 4 showed us that GSL is related to both the longitude and latitude of the station. It indicates that some spatial patterns might exist. Several years are randomly selected to identify the presence of spatial correlation of GSL. Most years presented similar patterns. 2009, showed here in Figure 4.1, is a typical case. Larger bubbles represent longer seasons length. Spatial patten can obviously be seen from the bubble plots. Stations at the southeast part of Kansas tend to have longer GSL.

In this chapter, a spatial-temporal model will be introduced and fitted for GSL. The model contains not only the time effect, but also the spatial.

4.1 Model Definition

In this section, we would like to give out the definition of spatial-temporal linear regression model(Reyes et. al 2012). Let $D_I = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_I\} \subset \mathbb{R}^d$ denote a spatial formation consisting of I sites $\boldsymbol{s}_i$, for $i = 1, \ldots, I$. Denote the response variable at site $\boldsymbol{s}_i \in D_I$ and time t as $y_{i,t} = y(\boldsymbol{s}_i, t)$, $i = 1, \ldots, I, t = 1, \ldots, T$. Meanwhile, denote a J-dimensional vector of

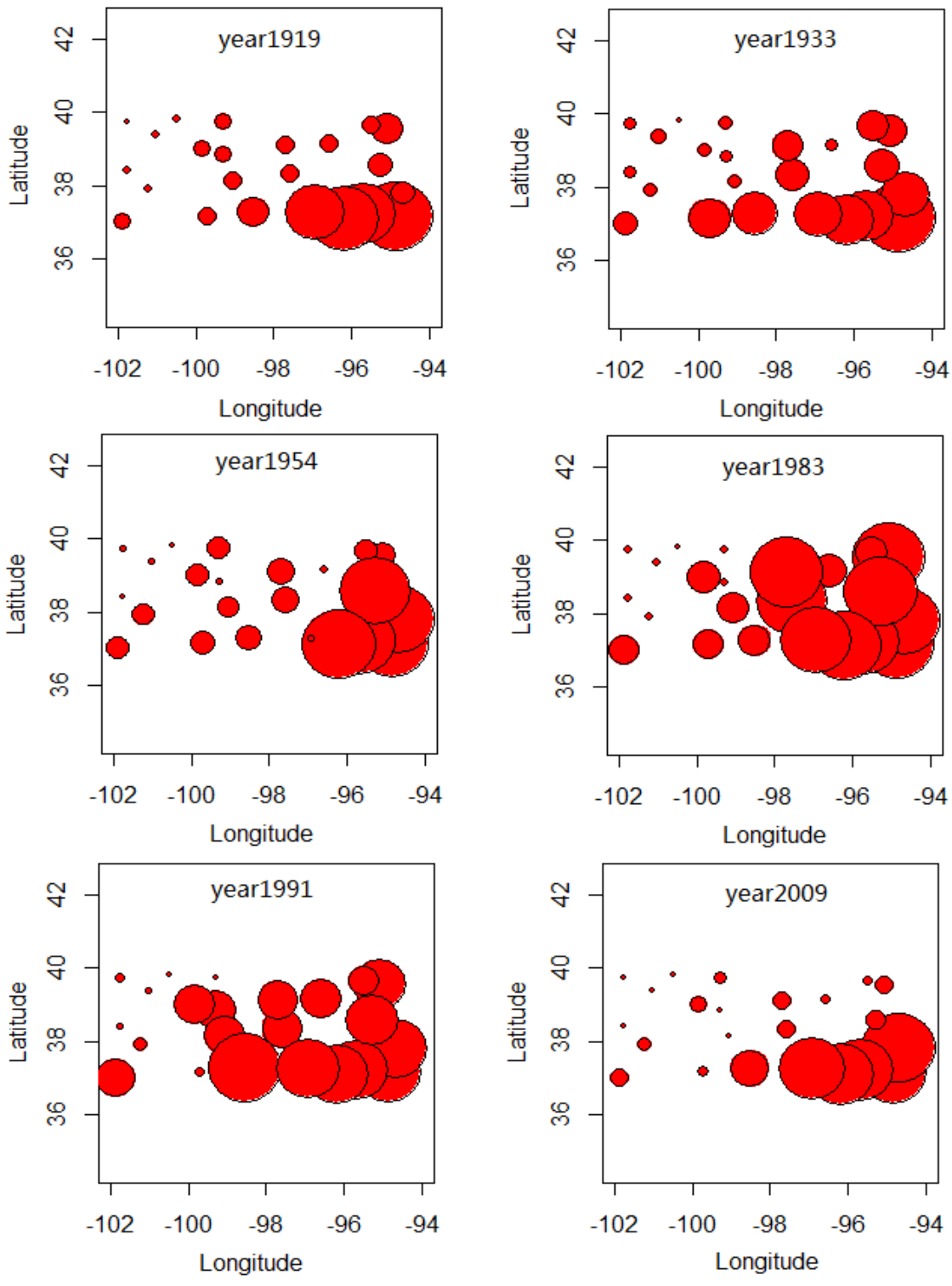

Figure 4.1: *Bubble plot of GSL from 6 randomly selected years. The bubble size is proportional to GSL.*

covariates at site $\boldsymbol{s}_i$ time t as $\boldsymbol{x}_{i,t} = (x_{1,i,t}, \ldots, x_{J,i,t})'$. Consider a linear regression model

$$y_{i,t} = \boldsymbol{x}'_{i,t}\boldsymbol{\beta} + \varepsilon_{i,t}, \tag{4.1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)'$ is a J-dimensional vector of regression coefficients. The error term is modeled by a spatial-temporal autoregressive model. Let

$$\boldsymbol{\varepsilon}_t = \sum_{l=0}^{L} \boldsymbol{C}_l \boldsymbol{\varepsilon}_{t-l} + \boldsymbol{\nu}_t, \tag{4.2}$$

where $\boldsymbol{\varepsilon}_t = (\varepsilon_{1,t}, \ldots, \varepsilon_{I,t})'$ denotes an I-dimensional vector of errors at time t for $t = 1, \ldots, T$, $L \geq 0$ is a pre-specified maximum time lag, and $\boldsymbol{C}_l$ for $l = 0, \ldots, L$ are $I \times I$ matrices consisting of $c_{i,i'}^{(l)}$ with $i, i' = 1, \ldots, I$. And $\boldsymbol{\nu}_t = (\nu_{1,t}, \ldots, \nu_{I,t})' \sim \text{iid}N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_I)$ consists of iid noise with mean 0 and variance component σ^2. Then the error term in the linear regression equation above can be described as

$$\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \boldsymbol{\Gamma}), \tag{4.3}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_{1,1}, \ldots, \varepsilon_{I,T})'$ denotes an N-dimensional vector of errors, and $\boldsymbol{\Gamma}$ is an $N \times N$ covariance matrix consisting of $\text{cov}(\varepsilon_{i,t}, \varepsilon_{i',t'})$, for $i, i' = 1, \ldots, I$, $t, t' = 1, \ldots, T$, and $N = IT$.

4.2 Defining the Neighborhood sturcture

For a spatial-temporal model, neighborhood structure also needs to be defined. For a given site i, we let $\mathcal{N}(i)$ be its neighborhood and let $\mathcal{N}(i) = \cup_{k=1}^{K} \mathcal{N}_k(i)$, where $\{\mathcal{N}_k(i) : k = 1, \ldots, K\}$ are neighborhoods that partition $\mathcal{N}(i)$, $i = 1, \ldots, I$ (Zhu et al. 2009). In this research, we divided the distances to site i into K intervals. The sites having distances to site i in the kth interval are considered to be the kth-order neighbors in $\mathcal{N}_k(i)$ of a given

site i.

We consider the parameterization introduced by Reyes and Zhu(Reyes et. al 2012 & Reyes 2010) for modeling spatial-temporal dependence:

$$\boldsymbol{C}_l = \sum_{k=0}^{K} \theta_{k,l} \boldsymbol{W}_{k,l}, \tag{4.4}$$

where $l = 0, \ldots, L$, $\theta_{k,l}$ is an unknown spatial-temporal coefficient, and $\boldsymbol{W}_{k,l} = [w_{i,i'}^{k,l}]_{i,i'=1}^{I}$ is an $I \times I$ matrix consisting of pre-specified spatial-temporal weights for the kth-order neighborhood and lth-order time lag, where $k = 0, \ldots, K$ and $l = 0, \ldots, L$. We assume that the weights are symmetric in the sense that $w_{i,i'}^{k,l} = w_{i',i}^{k,l}$ for all $i' \neq i$; $k = 1, \ldots, K$ and $l = 0, \ldots, L$. We set $\theta_{0,0} \equiv 0$ and $\boldsymbol{W}_{0,l} \equiv \boldsymbol{I}_I$ for $l \geq 1$ in order that at time lag $l = 0$, $\boldsymbol{C}_0 = \sum_{k=1}^{K} \theta_{k,0} \boldsymbol{W}_{k,0}$ features spatial autocorrelation among neighbors via spatial-only coefficients $\theta_{k,0}$ for $k = 1, \ldots, K$; and that at time lag $l \geq 1$, $\boldsymbol{C}_l = \theta_{0,l} \boldsymbol{I}_I + \sum_{k=1}^{K} \theta_{k,l} \boldsymbol{W}_{k,l}$ features spatial-temporal autocorrelation via temporal-only coefficients $\theta_{0,l}$ for $l = 1, \ldots, L$ and spatial-temporal coefficients $\theta_{k,l}$ for $k \geq 1$ and $l \geq 1$. For separable spatial-temporal model, we consider all $\boldsymbol{W}_{k,l}$, $k \geq 1$ to be zero. Because non-separable was much more complex and the results for our GSL data was not better. In this analysis, we used a separable spatial autoregressive (separable SAR) model which can be written as

$$y_{i,t} = x_{i,t}\beta + \epsilon_{i,t},$$

where $x_{i,t}$ are the covariates of each station. $\epsilon_{i,t}$ is the error term which can be described as

$$\epsilon_{i,t} = \Sigma_{l=1}{}^{L} \alpha_l \epsilon_{i,t-l} + \Sigma_{k=1}{}^{K} \theta_k W_k \epsilon_{i,t},$$

where α_l's are temporal autoregressive parameters and θ_k's are spatial autoregressive parameters.

Since we had a model selection algorithm, we were able to try different cases of neigh-

borhood structures to see which provided a better fit. Here, we have five different cases. In each case, we use circles with different radius to partition the neighborhood of stations. Because the shortest distance between these stations is 118 miles, the radius should be larger than 118 miles to make sure there is at least one neighbor for each station. For $CaseI$, we used 120 miles, 200 miles and 300 miles as radius to partition the neighborhood of stations into three parts. In $CaseI$ of Figure 4.2, we can see that for station i, its neighborhood is divided into 3 parts. The stations inside of the yellow-green circle with radius of 120 miles belong to the first order neighborhood, $N_1(i)$. Stations in the green part belong to the second order neighborhood $N_2(i)$. And stations in the blue part are in the third neighborhood, $N_3(i)$. $CaseII$ has exactly the same first 3 order neighborhoods as $CaseI$. The difference is that we combined neighbors inside of the green part and the yellow-green part as the forth neighborhood since the algorithm allowed multi-collinearity. For $CaseIII$, we used circles with 120 miles, 200 miles, 300 miles, 400 miles as radius to partition the neighborhood into four parts, while $CaseIV$ used 150 miles, 250 miles, 350 miles, 450 miles. $CaseV$ combined $CaseIII$ and $CaseIV$ together, and added neighbors within 200 miles as an additional neighborhood.

4.3 Model fitting

Let $\boldsymbol{\theta} = (\theta_{1,0}, \ldots, \theta_{K,0}, \ldots, \theta_{1,L}, \ldots, \theta_{K,L}, \theta_{0,1}, \ldots, \theta_{0,L})'$ denote an R-dimensional vector of spatial-temporal coefficients, where $R = (K+1)(L+1) - 1$. Henceforth, we replace the double index in $\theta_{k,l}$ with a single index θ_r, for $r = 1, \ldots, R$, except where double indexing aids interpretation. Let $\boldsymbol{\gamma} = (\boldsymbol{\theta}', \sigma^2)'$, we sometimes use $\boldsymbol{\Gamma}_{\boldsymbol{\gamma}}$ to emphasize the parameterization of $\boldsymbol{\Gamma}$ by $\boldsymbol{\gamma}$. Let $\boldsymbol{y} = (y_{1,1}, \ldots, y_{I,T})'$ denote an N-dimensional vector of response variables and let $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_J]$ denote an $N \times J$ design matrix, where $\boldsymbol{x}_j = (x_{j,1,1}, \ldots, x_{j,I,T})'$ denotes

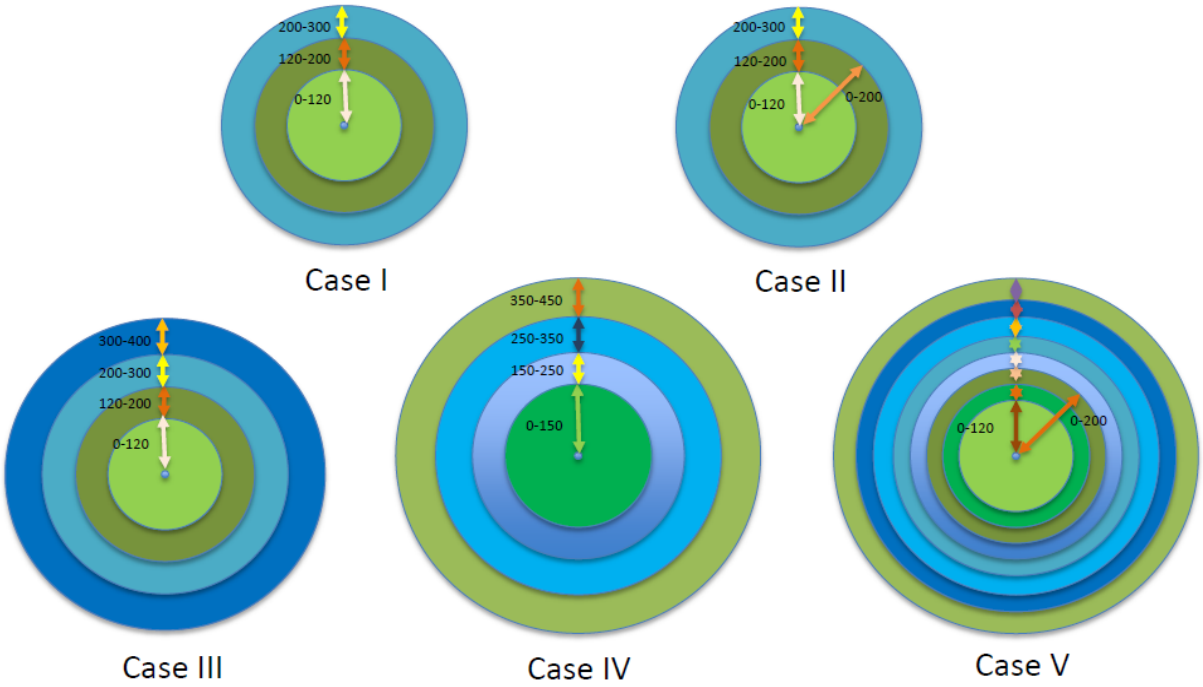

Figure 4.2: *Neighborhood partition plot*

an N-dimensional vector of the jth covariate with $j = 1, \ldots, J$. Thus, by (4.1) and (4.3),

$$\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\Gamma}_{\boldsymbol{\gamma}}). \tag{4.5}$$

For selection of a spatial-temporal dependence structure, we utilize the parameterization in (4.4) and determine which of the spatial-temporal coefficients are nonzero.

Let $\boldsymbol{\eta} = (\boldsymbol{\beta}', \boldsymbol{\gamma}')'$ denote a $(J+R+1)$-dimensional vector of model parameters consisting of both regression coefficients and spatial-temporal coefficients. Under (4.5), the log-likelihood function is

$$\begin{aligned} \log L(\boldsymbol{\eta}; \boldsymbol{y}, \boldsymbol{X}) &= \text{const} - (1/2)\log|\boldsymbol{\Gamma}_{\boldsymbol{\gamma}}| - (1/2)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{\Gamma}_{\boldsymbol{\gamma}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &\equiv \text{const} + \ell(\boldsymbol{\eta}). \end{aligned}$$

We let $\hat{\boldsymbol{\eta}}_{\text{MLE}} = \arg\max_{\boldsymbol{\eta}} \ell(\boldsymbol{\eta})$ denote the maximum likelihood estimates (MLE) of $\boldsymbol{\eta}$.

Consider the following penalized log-likelihood function

$$Q(\boldsymbol{\eta}) \quad = \quad \ell(\boldsymbol{\eta}) - N\sum_{j=1}^{J}\lambda_j|\beta_j| - N\sum_{r=1}^{R}\tau_r|\theta_r|, \tag{4.6}$$

where the last two terms are adaptive Lasso penalty on the coefficients, $\{\lambda_j\}_{j=1}^{J}$ are regularization parameters for the regression coefficients $\boldsymbol{\beta}$, and $\{\tau_r\}_{r=1}^{R}$ are regularization parameters for the spatial-temporal coefficients $\boldsymbol{\theta}$. We let $\hat{\boldsymbol{\eta}}_{\text{PMLE}} = \arg\max_{\boldsymbol{\eta}} Q(\boldsymbol{\eta})$ denote the penalized maximum likelihood estimates (PMLE) of $\boldsymbol{\eta}$. The details of the procedure are explained in Appendix B.

4.4 Results and conclusion

Using centered GSL values as response, spatial-temporal models for different neighborhood structures were fitted. The results are listed in Table 4.1-4.2. The covariates here include elevation, latitude, longitude and latitude:longitude.

Table 4.3, 4.4 list the results for all four cases by using standardized GSL as response.

From the results, it can be seen that different models can be fitted for different neighborhood structures. However, there are some common places among these models. First, according to Bayesian Information Criterion (BIC) (a criterion for model selection among a finite set of models), we picked $CaseIII$ as the optimal model. Second, the parameter for elevation are negative in all the models. An indication that GSL tends to be shorter when elevation gets higher. This conclusion is consistent with what we obtained by using the regular linear regression model. Third, the parameter for latitude is positive. Thus GSL tends to be longer for places with higher latitude in Kansas. Fourth, longitude were removed by the model selection procedure. Therefore, longitude was no longer significant when the spatial trend was considered. Finally, the standard deviation for latitude:longitude is very large, which indicates that the interaction of latitude and longitude was not significant.

	CaseI		CaseII		CaseIII		CaseIV		CaseV	
	Est	sd	Est	sd	Est	sd	Est	sd	Est	sd
Elevation β_1	-7.081	0.508	-7.283	0.520	-7.247	0.519	-7.024	0.492	-7.031	0.536
Latitude β_2	3.373	1.691	5.903	1.730	5.767	1.727	3.522	1.641	3.504	1.783
Longitude β_3	–	–	–	–	–	–	–	–	–	–
Lati:Long β_4	6.484	59.362	4.025	60.654	4.123	60.553	6.258	57.368	6.303	62.416
Spatial θ_{0-120}	–	–	-0.069	0.052	–	–			-0.042	0.056
$\theta_{120-200}$	0.117	0.025	–	–	0.101	0.025			–	–
$\theta_{200-300}$	0.156	0.032	0.150	0.032	0.131	0.032			0.064	0.071
θ_{0-200}			0.211	0.124					0.095	0.036
$\theta_{300-400}$					0.061	0.028			–	–
θ_{0-150}							–	–	0.158	0.145
$\theta_{150-250}$							0.088	0.031	–	–
$\theta_{250-350}$							0.085	0.031	–	–
$\theta_{350-450}$							0.094	0.026	0.051	0.060
Time Lag 1 α_1	0.041	0.020	0.041	0.020	0.023	0.020	–	–	0.017	0.020
2 α_2	-0.017	0.021	-0.017	0.020	–	–	–	–	–	–
3 α_3	–	–	–	–	–	–	–	–	–	–
4 α_4	–	–	–	–	–	–	–	–	–	–
5 α_5	-0.044	0.021	-0.044	0.021	-0.031	0.021	-0.012	0.021	-0.025	0.021
Variance σ^2	271.691	7.961	270.921	7.941	271.029	7.937	271.136	7.936	268.925	7.878
BIC	15584.41		15583.84		15577.93		15581.56		15583.22	

Table 4.1: *Fitted spatial temporal model for centered GSL, 5 time lags and different neighborhood structures: CaseI, 120 miles, 200 miles and 300 miles radius partition of the neighborhoods. CaseII same first 3 order neighborhoods as CaseI and the combined neighbors inside the 120&200 mile circle as the fourth neighborhood. CaseIII 120 miles, 200 miles, 300 miles, 400 miles radius partition of the neighborhood. CaseIV 150 miles, 250 miles, 350 miles, 450 miles radius. CaseV combined CaseIII and CaseIV, and added neighbors within 200 miles radius.*

	CaseI		CaseII		CaseIII		CaseIV		CaseV	
	Est	sd	Est	sd	Est	sd	Est	sd	Est	sd
Elevation β_1	-7.132	0.501	-7.121	0.517	-7.086	0.541	-7.101	0.540	-7.111	0.557
Latitude β_2	6.485	1.551	6.483	1.595	6.459	1.651	6.826	1.656	6.829	1.706
Longitude β_3	–	–	–	–	–	–	–	–	–	–
Lati:Long β_4	3.798	58.159	3.778	59.905	3.781	62.754	3.395	62.664	3.421	64.527
Spatial θ_{0-120}	–	–	-0.040	0.052	–	–			-0.045	0.056
$\theta_{120-200}$	0.112	0.025	–	–	0.110	0.025			–	–
$\theta_{200-300}$	0.144	0.032	0.137	0.032	0.136	0.032			0.073	0.070
θ_{0-200}			0.182	0.124					0.096	0.035
$\theta_{300-400}$					0.065	0.028			–	–
θ_{0-150}							0.018	0.025	0.166	0.144
$\theta_{150-250}$							0.107	0.030	–	–
$\theta_{250-350}$							0.094	0.030	–	–
$\theta_{350-450}$							0.101	0.025	0.052	0.059
Time Lag 1 α_1	0.021	0.020	0.023	0.020	0.031	0.020	0.020	0.020	0.024	0.020
2 α_2	–	–	–	–	–	–	–	–	–	–
3 α_3	–	–	–	–	–	–	–	–	–	–
4 α_4	–	–	0.000	0.021	–	–	–	–	–	–
5 α_5	-0.033	0.021	-0.035	0.021	-0.047	0.021	-0.042	0.021	-0.040	0.021
6 α_6	0.058	0.021	0.060	0.021	0.068	0.021	0.064	0.021	0.065	0.021
7 α_7	–	–	–	–	–	–	–	–	–	–
8 α_8	-0.097	0.021	-0.097	0.021	-0.102	0.021	-0.097	0.021	-0.097	0.021
9 α_9	-0.054	0.021	-0.055	0.021	-0.064	0.021	-0.060	0.021	-0.059	0.021
10 α_{10}	–	–	–	–	–	–	–	–	–	–
Variance σ^2	265.704	7.782	265.226	7.769	264.575	7.750	264.188	7.737	262.616	7.695
BIC	15550.480		15554.400		15545.400		15550.660		15547.620	

Table 4.2: *Fitted spatial temporal model for centered GSL, 10 time lags and different neighborhood structures: CaseI, 120 miles, 200 miles and 300 miles radius partition of the neighborhoods. CaseII same first 3 order neighborhoods as CaseI and the combined neighbors inside the 120&200 mile circle as the fourth neighborhood. CaseIII 120 miles, 200 miles, 300 miles, 400 miles radius partition of the neighborhood. CaseIV 150 miles, 250 miles, 350 miles, 450 miles radius. CaseV combined CaseIII and CaseIV, and added neighbors within 200 miles radius.*

	CaseI		CaseII		CaseIII		CaseIV		CaseV	
	Est	sd	Est	sd	Est	sd	Est	sd	Est	sd
Elevation β_1	-0.333	0.024	-0.332	0.024	-0.331	0.024	-0.330	0.023	-0.330	0.025
Latitude β_2	0.185	0.079	0.192	0.079	0.188	0.081	0.193	0.077	0.192	0.083
Longitude β_3	–	–	–	–	–	–	–	–	–	–
Lati:Long β_4	0.276	2.769	0.269	2.761	0.271	2.826	0.264	2.676	0.267	2.911
Spatial θ_{0-120}	–	–	-0.043	0.048	–	–			-0.042	0.056
$\theta_{120-200}$	0.117	0.025	–	–	0.102	0.025			–	–
$\theta_{200-300}$	0.156	0.032	0.075	0.023	0.131	0.032			0.064	0.071
θ_{0-200}			0.245	0.163					0.095	0.036
$\theta_{300-400}$					0.061	0.028			–	–
θ_{0-150}							–	–	0.158	0.145
$\theta_{150-250}$							0.088	0.031	–	–
$\theta_{250-350}$							0.085	0.031	–	–
$\theta_{350-450}$							0.094	0.026	0.050	0.060
Time Lag 1 α_1	0.041	0.020	0.024	0.020	0.023	0.020	–	–	0.017	0.020
2 α_2	-0.017	0.021	–	–	–	–	–	–	–	–
3 α_3	–	–	–	–	–	–	–	–	–	–
4 α_4	–	–	–	–	–	–	–	–	–	–
5 α_5	-0.044	0.021	-0.026	0.021	-0.031	0.021	-0.012	0.021	-0.025	0.021
Variance σ^2	0.591	0.017	0.591	0.017	0.590	0.017	0.590	0.017	0.585	0.017
BIC	1201.649		1205.505		1198.515		1198.750		1200.561	

Table 4.3: *Fitted spatial temporal model for standardized GSL, 5 time lags and different neighborhood structures: CaseI, 120 miles, 200 miles and 300 miles radius partition of the neighborhoods. CaseII same first 3 order neighborhoods as CaseI and the combined neighbors inside the 120&200 mile circle as the fourth neighborhood. CaseIII 120 miles, 200 miles, 300 miles, 400 miles radius partition of the neighborhood. CaseIV 150 miles, 250 miles, 350 miles, 450 miles radius. CaseV combined CaseIII and CaseIV, and added neighbors within 200 miles radius.*

	CaseI		CaseII		CaseIII		CaseIV		CaseV	
	Est	sd	Est	sd	Est	sd	Est	sd	Est	sd
Elevation β_1	-0.324	0.023	-0.324	0.024	-0.322	0.025	-0.321	0.025	-0.321	0.026
Latitude β_2	0.232	0.072	0.237	0.075	0.236	0.077	0.241	0.077	0.240	0.079
Longitude β_3	–	–	–	–	–	–	–	–	–	–
Lati:Long β_4	0.246	2.712	0.240	2.808	0.240	2.926	0.235	2.922	0.237	3.009
Spatial θ_{0-120}	–	–	-0.042	0.048	–	–			-0.044	0.056
$\theta_{120-200}$	0.112	0.025	–	–	0.109	0.025			–	–
$\theta_{200-300}$	0.144	0.032	0.075	0.023	0.136	0.032			0.072	0.070
θ_{0-200}			0.251	0.162					0.096	0.035
$\theta_{300-400}$					0.065	0.028			–	–
θ_{0-150}							0.018	0.025	0.165	0.144
$\theta_{150-250}$							0.106	0.030	–	–
$\theta_{250-350}$							0.095	0.030	–	–
$\theta_{350-450}$							0.100	0.025	0.053	0.059
Time Lag 1 α_1	0.021	0.020	0.028	0.020	0.032	0.020	0.021	0.020	0.024	0.020
2 α_2	–	–	–	–	–	–	–	–	–	–
3 α_3	–	–	–	–	–	–	–	–	–	–
4 α_4	–	–	–	–	–	–	–	–	–	–
5 α_5	-0.032	0.021	-0.038	0.021	-0.047	0.021	-0.042	0.021	-0.040	0.021
6 α_6	0.058	0.021	0.063	0.021	0.068	0.021	0.064	0.021	0.065	0.021
7 α_7	–	–	–	–	–	–	–	–	–	–
8 α_8	-0.097	0.021	-0.099	0.021	-0.102	0.021	-0.098	0.021	-0.097	0.021
9 α_9	-0.055	0.021	-0.058	0.021	-0.065	0.021	-0.061	0.021	-0.060	0.021
10 α_{10}	–	–	–	–	–	–	–	–	–	–
Variance σ^2	0.578	0.017	0.577	0.017	0.576	0.017	0.575	0.017	0.571	0.017
BIC	1171.015		1174.455		1165.851		1171.058		1167.997	

Table 4.4: *Fitted spatial temporal model for standardized GSL, 10 time lags and different neighborhood structures: CaseI, 120 miles, 200 miles and 300 miles radius partition of the neighborhoods. CaseII same first 3 order neighborhoods as CaseI and the combined neighbors inside the 120&200 mile circle as the fourth neighborhood. CaseIII 120 miles, 200 miles, 300 miles, 400 miles radius partition of the neighborhood. CaseIV 150 miles, 250 miles, 350 miles, 450 miles radius. CaseV combined CaseIII and CaseIV, and added neighbors within 200 miles radius.*

The results showed us that spatial trend does exist. It is consistent with the conclusion we obtained in Chapter 3 by using the regular linear regression model. Additionally, all the temporal parameters are very close to each other for different models. The association between the station and its neighbors within 200 miles are not significant for all spatial-temporal models. While the association between the station and its neighbors further than 200 miles are significant. The time lag 1, 5, 6, 8, 9 were significant for all the models. This matches with the existence of higher order temporal correlation revealed in Chapter 2.

The Bayesian information criterion (BIC) is a criterion for model selection among a finite set of models. According to the BIC values of the fitted models, we suggest $CaseIII$ as the optimal model.

Chapter 5

Discussion

Frost indices affect not only natural and managed ecosystems, but also a variety of human activities. In particular for agricultural activities, we might be interested in using such indices as guidelines for planning. Growing season length is one of the most important frost indices. It is very useful in determining crop cycle lengths and calendars under average conditions. In this report, data collected from late 1800's to 2009 at 23 long term stations across Kansas were analyzed.

Firstly, we fitted time series models to GSL of each station. From the fitted time series models, it could be seen that different stations had different models with different complexity. Stations with higher elevations tended to have more complicated models and stations with lower elevations tent to have relatively simpler models. This implied that GSL could be associated with elevation and/or location.

Since there might be some association between GSL and elevation and/or location, we applied regular linear regression on year, elevation, latitude and longitude to all stations GSL data combined. The obtained linear regression model showed us that the GSL did have a significant association with elevation. The sign of the parameter for elevation in the selected model was negative, which indicated that stations with higher elevations tent to have shorter GSL than stations with lower elevations. The coefficient of latitude was

positive and the coefficient for longitude was negative. So it tended to have longer growing season length for larger latitude (to the north) and smaller longitude (to the west). By checking the residuals, we found that spatial-temporal trend did exist for GSL and it wasn't accounted by the inclusion of latitude, longitude, elevation and year.

Finally, to explain the spatial pattern better, we fitted spatial-temporal linear regression model of GSL for different neighborhood structures. According to the BIC values of the fitted models, the optimal model was selected. The parameter for elevation in the model were all negative. It meant that GSL tended to be shorter when elevation got higher. This conclusion was consistent with what we obtained by using the linear regression model. The parameter for latitude was positive. Thus, GSL tended to be longer for places with higher latitude in Kansas. Longitude were removed by the model selection procedure, which indicated that longitude were no longer significant when the spatial trend was considered. The standard deviation for latitude:longitude were very large, implying that the interaction of latitude and longitude was not significant.

All the regression models fitted in this report were linear models. For future work, we can try non-linear spatial-temporal model to see if it is more flexible to explain GSL trend. To verify the model, we can fit the spatial-temporal model for GSL in other state to see if there are similarities.

Bibliography

[1] USGCRP (2009). Global Climate Change Impacts in the United States . Karl, T.R., J.M. Melillo, and T.C. Peterson (eds.). United States Global Change Research Program. Cambridge University Press, New York, NY, USA.

[2] U.S. Census of Agriculture, 2012

[3] Hatfield, J., 2012: Agriculture in the Midwest. In: U.S. National Climate Assessment Midwest Technical Input Report. J. Winkler, J. Andresen, J. Hatfield, D. Bidwell, and D. Brown, coordinators. Available from the Great Lakes Integrated Sciences and Assessments (GLISA) Center, [1].

[4] Kansaspedia, [online]. Available: https://www.kshs.org/kansapedia/agriculture-in-kansas/14188.

[5] Anandhi, Aavudai, et al(2013). "Long-term spatial and temporal trends in frost indices in Kansas, USA." Climatic Change 120.1-2 (2013): 169-181.

[6] Meehl, G. A., Claudia Tebaldi, and Doug Nychka. "Changes in frost days in simulations of twentyfirst century climate." Climate Dynamics 23.5 (2004): 495-511.

[7] Cressie, Noel AC, and Noel A. Cassie(1993). Statistics for spatial data. Vol. 900. New York: Wiley, 1993.

[8] Reyes, Perla E., Jun Zhu, and Brian H. Aukema(2012). "Selection of Spatial-Temporal Lattice Models: Assessing the Impact of Climate Conditions on a Mountain Pine Beetle Outbreak." Journal of agricultural, biological, and environmental statistics 17.3 (2012): 508-525.

[9] Reyes, Perla E.(2010). "Selection of spatial and spatial-temporal linear models for lattice data" PhD dissertation at University of Wisconsin-Madison(2010).

[10] Shumway, Robert H., David S. Stoffer, and David S. Stoffer. Time series analysis and its applications. Vol. 3. New York: Springer, 2000.

[11] Zhu, Zhengyuan, and Yufeng Liu. "Estimating spatial covariance using penalised likelihood with weighted L 1 penalty." Journal of Nonparametric Statistics 21.7 (2009): 925-942.

[12] Tibshirani, R. (1996), Regression Shrinkage and Selection via the Lasso, Journal of the Royal Statistical Society, Series B, 58, 267288.

[13] Box, George EP, Gwilym M. Jenkins, and Gregory C. Reinsel. Time series analysis: forecasting and control. John Wiley & Sons, 2013.

Appendix A

Time series analysis 18 remaining stations

In chapter 2, we only listed the time series plots, prediction plots and validation plots for five selected stations. Plots for the rest 18 stations are listed here from Figure A.1-A.9.

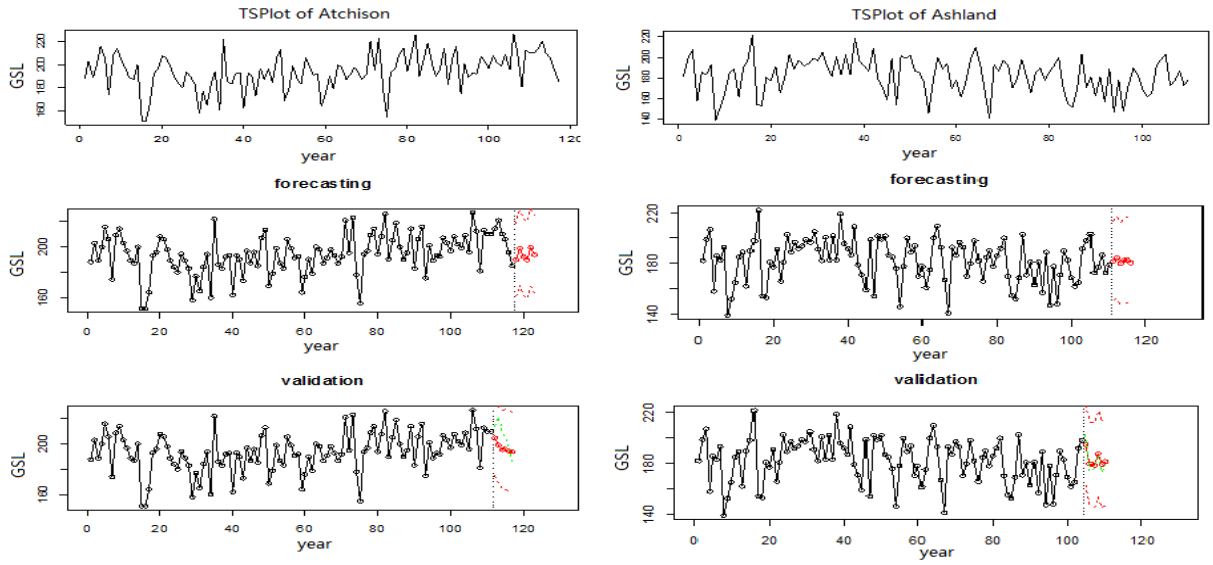

Figure A.1: *Time series plot forcasting and validation plots Atchison and Ashland stations*

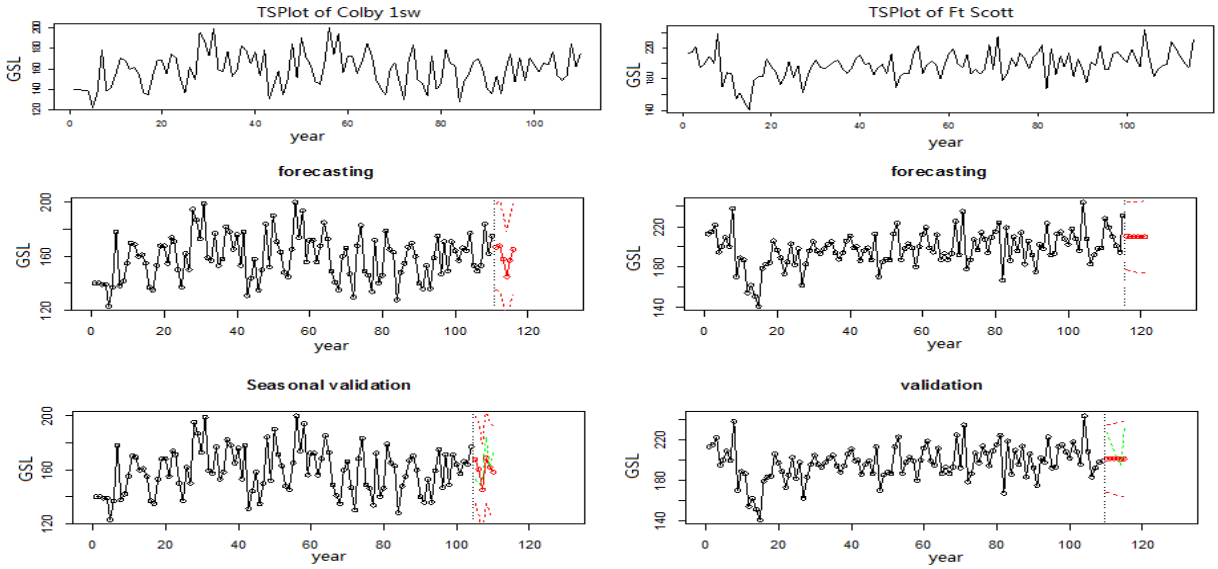

Figure A.2: *Time series plot forcasting and validation plots Colby 1sw and Ft Scott stations*

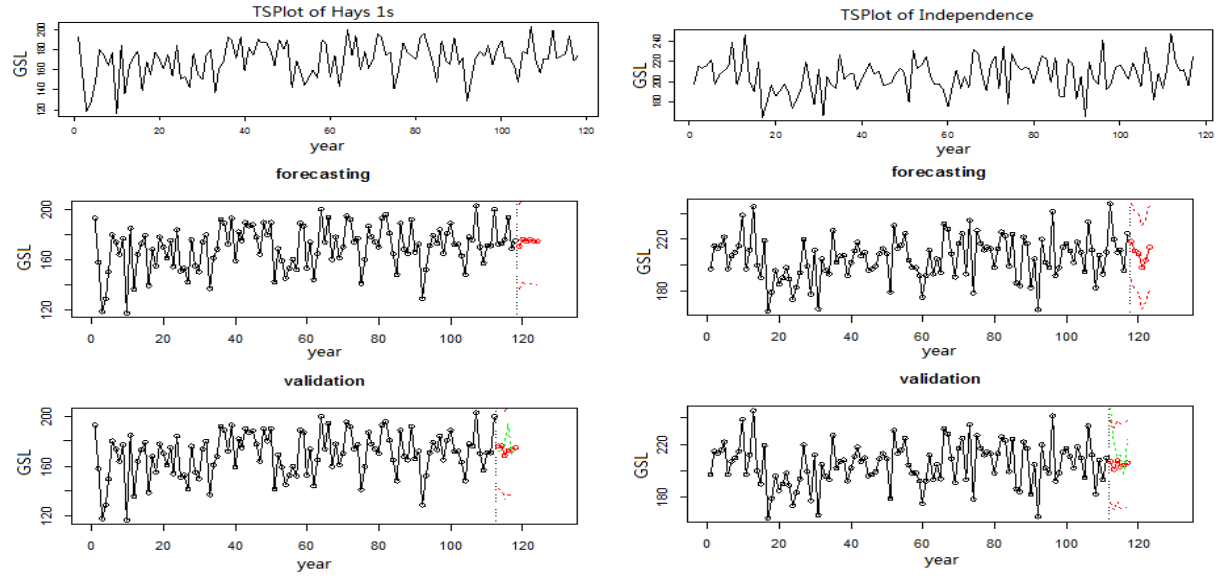

Figure A.3: *Time series plot forcasting and validation plots Hays 1s and Independence stations*

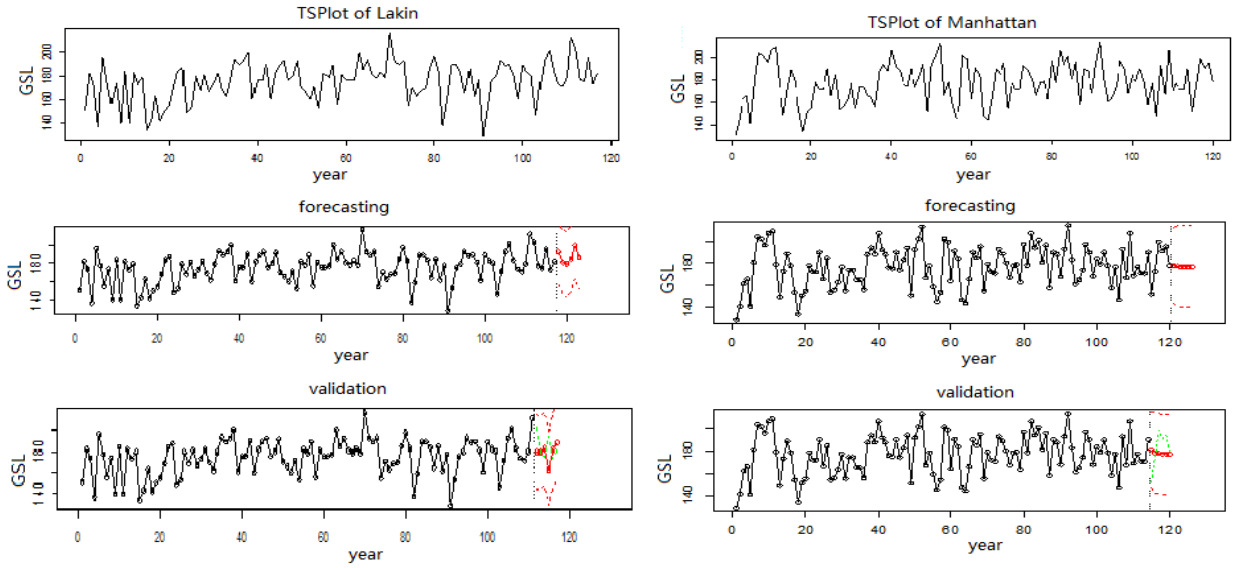

Figure A.4: *Time series plot forcasting and validation plots Lakin and Manhattan stations*

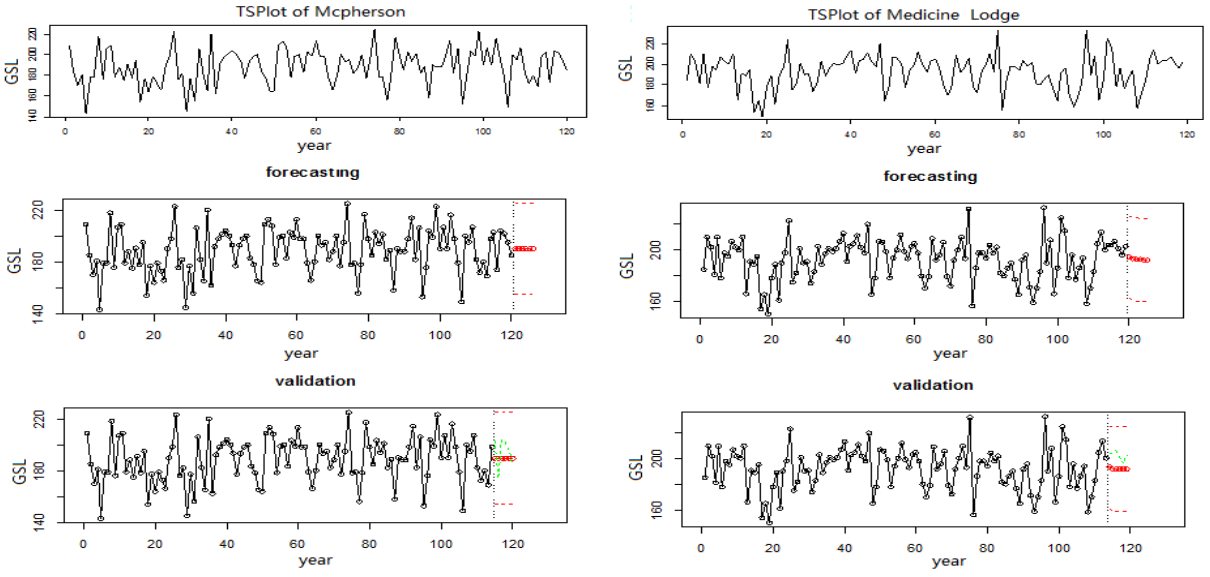

Figure A.5: *Time series plot forcasting and validation plots Mcpherson and Medicine Lodge stations*

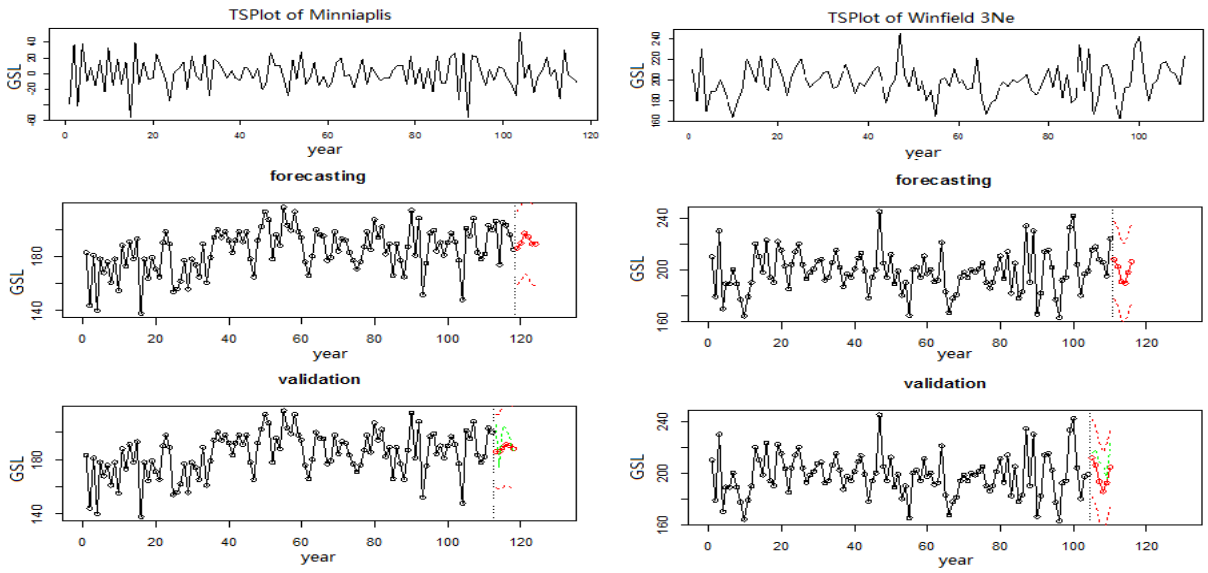

Figure A.6: *Time series plot forcasting and validation plots Minniapolis and Winfield 3Ne stations*

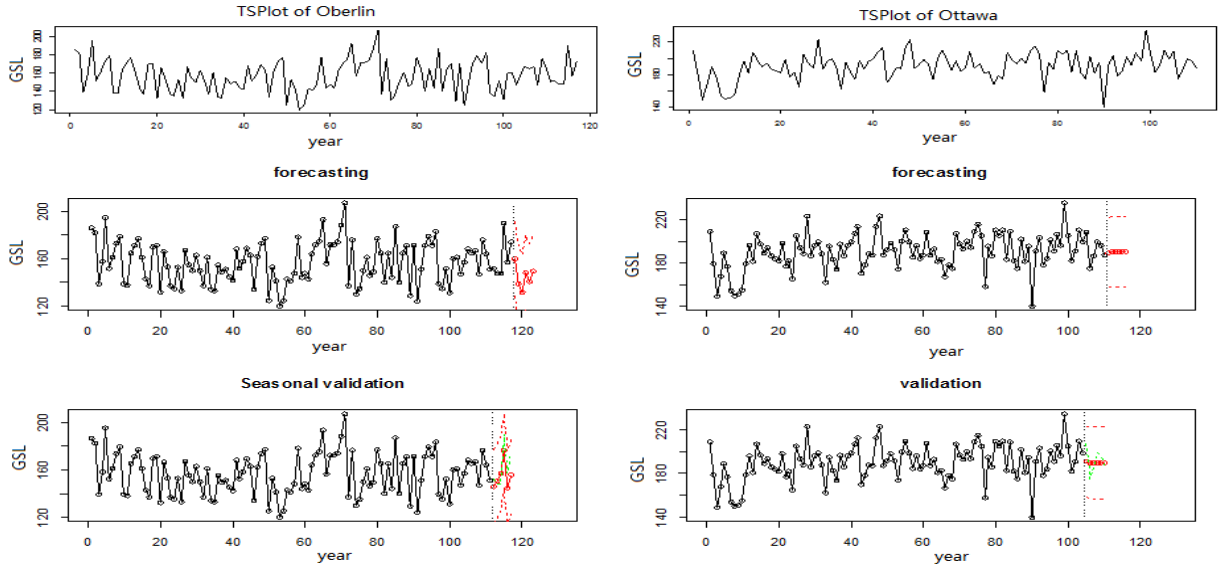

Figure A.7: *Time series plot forcasting and validation plots Oberlin and Ottawa stations*

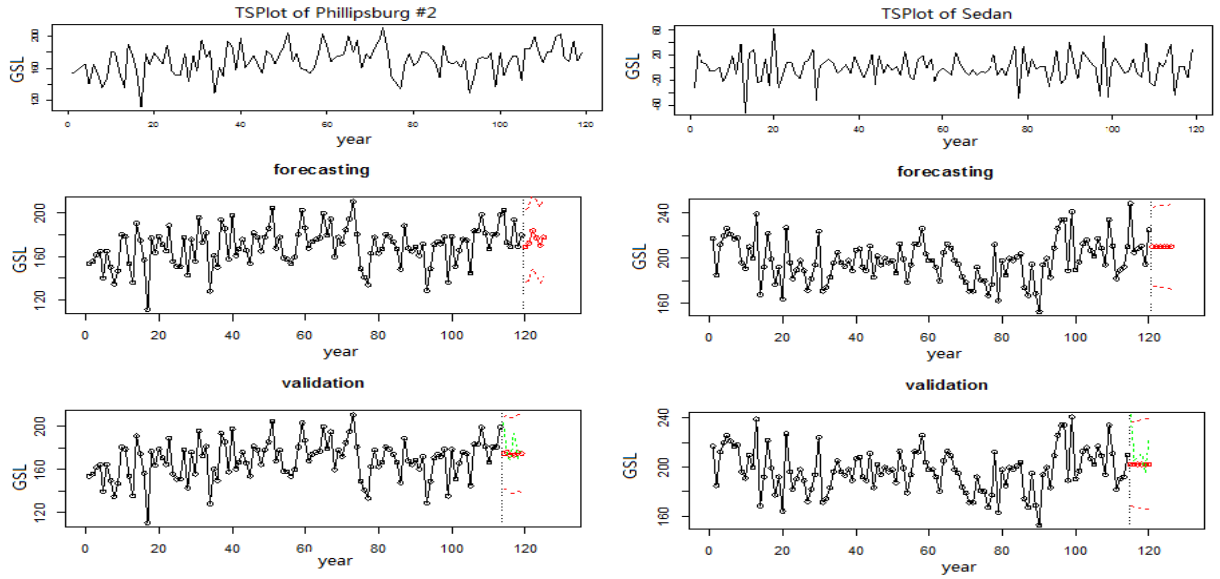

Figure A.8: *Time series plot forcasting and validation plots Phillipsburg and Sedan stations*

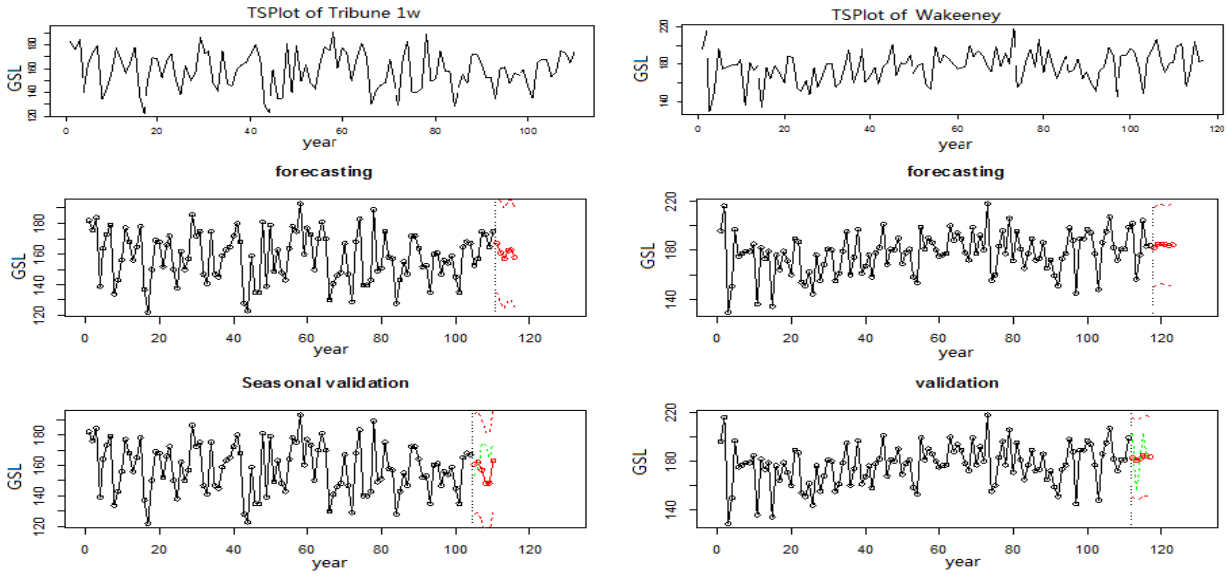

Figure A.9: *Time series plot forcasting and validation plots Tribune 1w and Wakeeney stations*

Appendix B

Some details of penalized likelihood method

In this Appendix, some details of penalized likelihood method (Reyes 2012, Tibshirani 1996) are introduced. Let $\hat{\boldsymbol{\eta}}^{(0)} = (\hat{\boldsymbol{\beta}}^{(0)'}, \hat{\boldsymbol{\gamma}}^{(0)'})'$ denote an initial value of $\boldsymbol{\eta}$, which is set to the MLE $\hat{\boldsymbol{\eta}}_{\text{MLE}}$. Given that $\boldsymbol{\eta} \approx \hat{\boldsymbol{\eta}}^{(0)}$, we approximate the penalized log-likelihood function (4.6) up to a constant by

$$\begin{aligned} Q^*(\boldsymbol{\eta}) &= (\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}^{(0)})' \frac{\partial \ell(\hat{\boldsymbol{\eta}}^{(0)})}{\partial \boldsymbol{\eta}} - (1/2)(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}^{(0)})' \mathcal{I}(\hat{\boldsymbol{\eta}}^{(0)})(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}^{(0)}) \\ &\quad - N \sum_{j=1}^{J} \lambda_j |\beta_j| - N \sum_{r=1}^{R} \tau_r |\theta_r|, \end{aligned} \tag{B.1}$$

where $\mathcal{I}(\boldsymbol{\eta}) = \mathrm{E}_{\boldsymbol{\eta}}\Big\{ - \frac{\partial^2 \ell(\boldsymbol{\eta})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}'} \Big\}$ is an expected information matrix (Zhu et al. 2009). We propose to approximate $\hat{\boldsymbol{\eta}}_{\text{PMLE}}$ by

$$\hat{\boldsymbol{\eta}}^{(1)} = \arg\max_{\boldsymbol{\eta}} \Big\{ Q^*(\boldsymbol{\eta}) \Big\}. \tag{B.2}$$

Since the expected information matrix is block diagonal with $\mathcal{I}(\boldsymbol{\eta}) = \text{diag}\{\mathcal{I}(\boldsymbol{\beta}), \mathcal{I}(\boldsymbol{\gamma})\}$,

we obtain $\hat{\boldsymbol{\beta}}^{(1)}$ and $\hat{\boldsymbol{\gamma}}^{(1)}$ separately. That is,

$$\hat{\boldsymbol{\beta}}^{(1)} = \arg\min_{\boldsymbol{\beta}} \Big\{ -(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}^{(0)})'\frac{\partial \ell(\hat{\boldsymbol{\eta}}^{(0)})}{\partial \boldsymbol{\beta}} + (1/2)(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}^{(0)})'\mathcal{I}(\hat{\boldsymbol{\beta}}^{(0)})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}^{(0)}) + N\sum_{j=1}^{J}\lambda_j|\beta_j| \Big\}. \tag{B.3}$$

It can be shown that the solution of (B.3) can be attained equivalently by

$$\hat{\boldsymbol{\beta}}^{*(1)} = \arg\min_{\boldsymbol{\beta}^*} \Big\{ (1/2)(\boldsymbol{y}^* - \boldsymbol{X}^*\boldsymbol{\beta}^*)'(\boldsymbol{y}^* - \boldsymbol{X}^*\boldsymbol{\beta}^*) + N\sum_{j=1}^{J}|\beta_j^*| \Big\}, \tag{B.4}$$

where $\boldsymbol{y}^* = (\boldsymbol{A}^{-1})'\{\frac{\partial \ell(\hat{\boldsymbol{\eta}}^{(0)})}{\partial \boldsymbol{\beta}} + \mathcal{I}(\hat{\boldsymbol{\beta}}^{(0)})'\hat{\boldsymbol{\beta}}^{(0)}\}$, $\boldsymbol{X}^* = \boldsymbol{A}\text{diag}\{\lambda_j^{-1}\}_{j=1}^J$, $\boldsymbol{\beta}^* = \text{diag}\{\lambda_j\}_{j=1}^J\boldsymbol{\beta}$, and $\mathcal{I}(\hat{\boldsymbol{\beta}}^{(0)}) = \boldsymbol{A}'\boldsymbol{A}$. Hence, $\hat{\boldsymbol{\beta}}^{(1)} = \text{diag}\{\lambda_j^{-1}\}_{j=1}^J\hat{\boldsymbol{\beta}}^{*(1)}$.

Next,

$$\hat{\boldsymbol{\gamma}}^{(1)} = \arg\min_{\boldsymbol{\gamma}} \Big\{ -(\boldsymbol{\gamma}-\hat{\boldsymbol{\gamma}}^{(0)})'\frac{\partial \ell(\hat{\boldsymbol{\eta}}^{(0)})}{\partial \boldsymbol{\gamma}} + (1/2)(\boldsymbol{\gamma}-\hat{\boldsymbol{\gamma}}^{(0)})'\mathcal{I}(\hat{\boldsymbol{\gamma}}^{(0)})(\boldsymbol{\gamma}-\hat{\boldsymbol{\gamma}}^{(0)}) + N\sum_{r=1}^{R}\tau_r|\theta_r| \Big\}. \tag{B.5}$$

Given that σ^2 is not subject to any penalty, we let

$$\boldsymbol{X}_r^{**} = \tau_r^{-1}(\boldsymbol{B}_r - c_r\boldsymbol{B}_{R+1}), \quad r = 1,\ldots,R, \quad \text{and} \quad \boldsymbol{X}_{R+1}^{**} = \boldsymbol{B}_{R+1},$$

where $c_r = \boldsymbol{B}_{R+1}'\boldsymbol{B}_r/\boldsymbol{B}_{R+1}'\boldsymbol{B}_{R+1}$, for $r = 1,\ldots,R$, and $\mathcal{I}(\hat{\boldsymbol{\gamma}}^{(0)}) = \boldsymbol{B}'\boldsymbol{B}$. It follows that $\boldsymbol{X}_{R+1}^{**'}\boldsymbol{X}_r^{**} = \boldsymbol{0}$ for $r = 1,\ldots,R$. Let $\boldsymbol{y}^{**} = (\boldsymbol{B}^{-1})'\{\frac{\partial \ell(\hat{\boldsymbol{\eta}}^{(0)})}{\partial \boldsymbol{\gamma}} + \mathcal{I}(\hat{\boldsymbol{\gamma}}^{(0)})'\hat{\boldsymbol{\gamma}}^{(0)}\}$. It can be shown that the solution of σ^2 in (B.5) has a closed form $(\hat{\sigma}^{*2})^{(1)} = \boldsymbol{X}_{R+1}^{**'}\boldsymbol{y}^{**}/\boldsymbol{X}_{R+1}^{**'}\boldsymbol{X}_{R+1}^{**}$, where $\sigma^{*2} = \sum_{r=1}^R c_r\theta_r + \sigma^2$. Furthermore,

$$\hat{\boldsymbol{\theta}}^{*(1)} = \arg\min_{\boldsymbol{\theta}^*} \Big\{ (1/2)(\boldsymbol{y}^{**} - \boldsymbol{X}^{**}\boldsymbol{\theta}^*)'(\boldsymbol{y}^{**} - \boldsymbol{X}^{**}\boldsymbol{\theta}^*) + N\sum_{r=1}^{R}|\theta_r^*| \Big\}, \tag{B.6}$$

where $\boldsymbol{X}^{**} = [\boldsymbol{X}_1^{**}, \ldots, \boldsymbol{X}_R^{**}]$ and $\boldsymbol{\theta}^* = \text{diag}\{\tau_r\}_{r=1}^R \boldsymbol{\theta}$. Hence, $\hat{\boldsymbol{\theta}}^{(1)} = \text{diag}\{\tau_r^{-1}\}_{r=1}^R \hat{\boldsymbol{\theta}}^{*(1)}$ and $(\hat{\sigma}^2)^{(1)} = (\hat{\sigma}^{*2})^{(1)} - \sum_{r=1}^R c_r \hat{\theta}_r^{(1)}$.

Let $\hat{\boldsymbol{\eta}}_{\text{APMLE}} = \hat{\boldsymbol{\eta}}^{(1)}$ denote the approximate penalized maximum likelihood estimates (APMLE) of $\boldsymbol{\eta}$, where $\hat{\boldsymbol{\eta}}^{(1)} = (\hat{\boldsymbol{\beta}}^{(1)'}, \hat{\boldsymbol{\gamma}}^{(1)'})'$. Equations (B.4) and (B.6) can be solved by a LARS algorithm and thus, the computation is efficient. Although Equation (B.2) can be iterated until convergence, a one-step solution is preferred here because it is computationally efficient and the estimates still possess desirable asymptotic properties, as is explained in Reyes et al..