Statistical Modeling and Analysis of Bivariate Spatial-Temporal Data with the Application to Stream Temperature Study

Han Li

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Statistics

Eric P. Smith, Co-Chair
Xinwei Deng, Co-Chair
Hongxiao Zhu
C. Andrew Dolloff

October 8th, 2014
Blacksburg, Virginia

KEYWORDS: Steam Temperature, Varying Coefficient Model, Functional Data Clustering, Missing Data Imputation

Statistical Modeling and Analysis of Bivariate Spatial-Temporal Data with the Application to Stream Temperature Study

Han Li

ABSTRACT

Water temperature is a critical factor for the quality and biological condition of streams. Among various factors affecting stream water temperature, air temperature is one of the most important factors related to water temperature. To appropriately quantify the relationship between water and air temperatures over a large geographic region, it is important to accommodate the spatial and temporal information of the steam temperature. In this dissertation, I devote effort to several statistical modeling techniques for analyzing bivariate spatial-temporal data in a stream temperature study.

In the first part, I focus our analysis on the individual stream. A time varying coefficient model (VCM) is used to study the relationship between air temperature and water temperature for each stream. The time varying coefficient model enables dynamic modeling of the relationship, and therefore can be used to enhance the understanding of water and air temperature relationships. The proposed model is applied to 10 streams in Maryland, West Virginia, Virginia, North Carolina and Georgia using daily maximum temperatures. The VCM approach increases the prediction accuracy by more than 50% compared to the simple linear regression model and the nonlinear logistic model.

The VCM that describes the relationship between water and air temperatures for each stream is represented by slope and intercept curves from the fitted model. In the second part, I consider water and air temperatures for different streams that are spatial correlated. I focus on clustering multiple streams by using intercept and slope curves estimated from the VCM. Spatial information is incorporated to make clustering results geographically meaningful. I further propose a weighted distance as a dissimilarity measure for streams, which provides a flexible framework to interpret the clustering results under different weights. Real data analysis shows that streams in same cluster share similar geographic features such as solar radiation, percent forest and elevation.

In the third part, I develop a spatial-temporal VCM (STVCM) to deal with missing data. The STVCM takes both spatial and temporal variation of water temperature into account. I develop a novel estimation method that emphasizes the time effect and treats the space effect as a varying coefficient for the time effect. A simulation study shows that the performance of the STVCM on missing data imputation is better than several existing methods such as the neural network and the Gaussian process. The STVCM is also applied to all 156 streams in this study to obtain a complete data record.

Dedicated to my parents.

Acknowledgments

Three years ago, I would never think of a moment writing this chapter. The way to finishing this dissertation is full of doubt about myself. I still cannot believe what happened when Dr. Smith told me I passed the final examination. This dissertation will not be completed without the help of my advisers, Dr. Eric P. Smith and Dr. Xinwei Deng. I am so grateful that they not only provide insightful guide, consistent encouragement or even criticism, but also, and more importantly, they always believe in my competence. In the year of 2012, when Dr. Deng and I were talking about future of statistics, I was informed that Dr. Smith has a project that may lead to PhD dissertation. Since then, I have been working on this research problem for three years and enjoy it. After the talk with the Climate Corporation last week, I realize that the spatial-temporal research on weather might be future of statistics.

Dr. Deng is one of the smartest person I have met. No matter how difficult the problems we faced, he can always figure out a way to solve them. He is more like a strategist in my dissertation. His intelligence makes me feel comfortable to propose any methods and try any new ideas. He is also a forceful man. I still remember the time when we needed to define the distance in the variogram, he taught me never to give things up. I really hope my dissertation lives up to his standards.

Dr. Smith has an easy going demeanor. He is the first person to teach me how to make things work. So most of the time we focus on real cases and interpreting the results. I think his philosophy is that it is not how complicated the model is; it is how good the model is. Due to this logic, our first paper got successfully published. To me, Dr. Smith is a supervisor in life than in school work. The lessons I learned how to reply to the reviewers, how to answer tough questions and how to overcome troubles will be helpful in my life time. I think I got most of his humors in meetings but he might not believe it.

I also want to thank other professors in my committee and in other departments. Dr. Hongxiao Zhu is very pleasant to speak with and she treats me as her own PhD student when we talk. Her suggestions on the functional data analysis and the software packages she provided are really helpful and saved me much time. I appreciate her always being available for my appointments and questions. Dr. Andy C. Dolloff gave the biological meanings for our statistical models which makes me feel confident in our results. I would like to express my appreciation to Dr. Dolloff. He has a very busy schedule but is still willing to take time

to introduce the fishery background to me. I thank Prof. Mike Kender and Dr. Vijay Singal for their help on fixed income and equity researches.

During the study in Virginia Tech, I met many friends who make my life in Blacksburg joyful. Zhen Liu, Peng Zhang and Will Hozey are all the greatest roommates I ever have. I enjoyed the time with Joe Bartlett and Josh Freivogel for all the new thing I have tried. Penny Park, Ying Powers, Eric Fu, Chris Wang and Caroline Yang are always supportive whenever I need them.

Han Li
Oct 13th, 2014

Contents

Chapter 1

Introduction

1.1 Statement of the Problem and Background

This dissertation aims to model the relationship between water and air temperatures across time and space. Specifically, the interest lies in the sensitivity of water temperature to the changes in air temperature and how this sensitivity varies across time and space. Other problems of interests are using air temperature to predict water temperature and organizing streams into management units based on the water-air relationship. The proposed methods used in this dissertation can also be applied to bivariate time series in other areas such as finance (S&P500 index vs S&P500 future index), marketing (sales vs commercial costs) and social networks (number of tweets vs number of followers).

Water temperature is a critical component of hydrologic systems (Keleher and Rahel, 1996; Caissie, 2006) and can be a determining factor in water quality and biological condition. Fish and other aquatic organisms are sensitive to extremes in temperature, with extremes affecting food sources, survival and the range of organisms. Brook trout, for example, prefers

cooler water found in high elevation streams, and temperatures greater than $21°C$ are considered as highly stressful to the health of trout (Meisner, 1990; Beitinger *et al.*, 2000). Growth, reproduction, migration and food availability are all affected by water temperature (see Caissie (2006) for a review). Factors critical to the health of aquatic systems, such as dissolved oxygen, are also dependent on temperature. Warm water tends to hold less dissolved oxygen than cold water, and concentrations of dissolved oxygen tend to be lower in the summer and fall. Effects of temperature change therefore can lead to significant biological effects. Stefan *et al.* (2001), for example, predict that increased water temperature will result in increased summertime killing of fish in lakes. Flebbe *et al.* (2006) expect that changes in temperature will result in dramatic changes in the habitat for brook trout (see also Keleher and Rahel (1996); Minns *et al.* (1995)).

Models describing the change in water temperature associated with changes in air temperature are therefore critical in determining the effects of temperature change on aquatic systems. Hence measuring and visualizing sensitivity of water temperature to changes in air temperature is an important goal. Various models have been proposed to predict water temperature using air temperature data at different time scales. The choice of a model depends on the availability of information on factors affecting water temperature, the objective of the study, temporal and spatial aspects of the data, as well as the time step and duration of measurement (Mohseni *et al.*, 1998; Caissie, 2006; Mayer, 2012). In this dissertation, I develop models for quantifying the relationship between daily maximum air and water temperatures and consequently use daily maximum air temperature to predict daily maximum water temperature. Daily maximum temperature is selected over weekly average temperature as the variable of interest because the maximum water temperature has been linked to loss of brook trout, a temperature sensitive species (Trumbo *et al.*, 2010).

The management of trout streams in terms of fish habitat and high water temperature risk

will be more effective if streams can be organized into management units. A key question to address is whether or not there are clusters of streams with similar profiles for air and water temperatures. This objective calls for clustering streams based on water and air temperature relationships. If streams within the same cluster have similar profiles of risk, then investments may be made to better manage streams tailored to the water-air relationship (Mayer, 2012). Clustering streams is also valuable for other objectives. For example, there are many climate and landscape factors affecting water and air temperature relationship in streams such as solar radiation and forest percent (Chen *et al.*, 1998). Those factors tend to be distinct for each stream but may show some similarity in the same cluster.

Climate change is one of most important stressors for fish populations (Flebbe *et al.*, 2006). Evidence indicates that climate has changed for the last decade (Alexander, 2013). Although there are observed changes in atmosphere, ocean, cryosphere, sea level, carbon and other biogeochemical cycles on global level, it is not clear how climate change affects local hydrologic systems in eastern US. The Eastern Brook trout Joint Venture (EBTJV) conducted a study to assess the relationship between trout and water temperature at over 5000 sites in the eastern United States (EBTJV, 2006).

As part of EBTJV study, paired (air and water) thermographs (HOBO Watertemp Pro v2; accuracy $0.2°C$; drift < 0.1 annually (Onset Computer Corporation, 2009)) were placed at the pour point of randomly selected 156 stream catchments in southeast USA since the year 2009. (The pour point or outlet of the watershed is the point in the watershed that all water flows through.) The location of streams are shown in Figure 1.1. The stream catchments were selected from a statistical population of over 1000 catchments known to support brook trout. A stratified approach was used based on information on area of the catchment, elevation, forest cover and solar radiation. Additional details on site characteristics, study design and sampling were given in Trumbo *et al.* (2010). All thermographs were set to

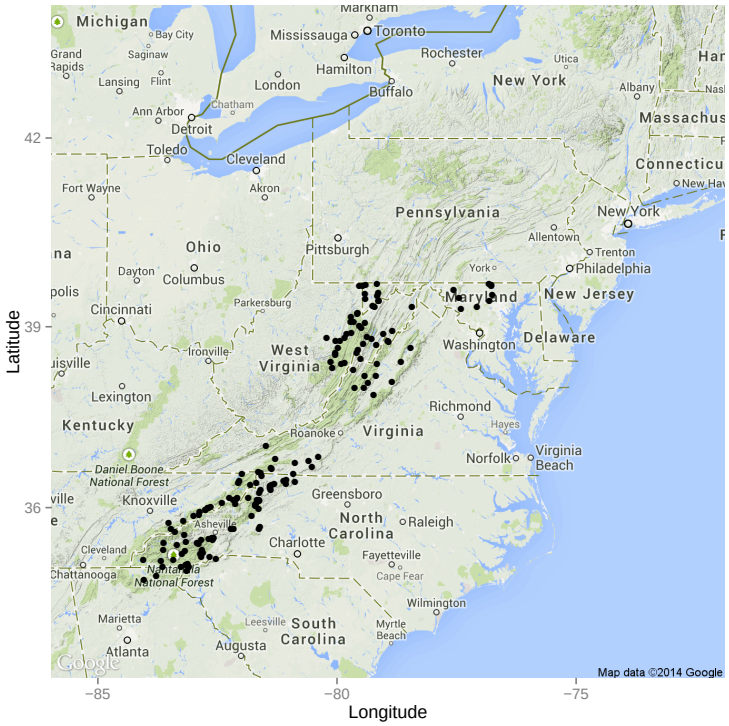

Figure 1.1: Location of 156 streams (black dots are the locations).

record every 30 minutes (Dunham *et al.*, 2005; Huff *et al.*, 2005) since 2009. Thermographs were calibrated before and after deployment following methods summarized in Dunham *et al.* (2005). Because stream channels may become dry during summer low flow periods, thermographs used to record water temperatures were placed near the deepest location in the stream segment when possible (Lisle, 1987). A shield was used to reduce direct solar radiation heating on air temperature thermographs (Dunham *et al.*, 2005; Trumbo *et al.*, 2012).

The raw air and water temperature data were screened for identifying outliers and other oddities resulting from thermograph malfunctions, launch/recording interval errors, or potentially dry stream beds. Preliminary analyses were also conducted to ensure data quality. Scatterplots of water and air temperatures were drawn to evaluate the joint relationship and to look for irregularities in the daily maximum air and water temperature. Index plots of

lagged one-day differences were also used to identify oddities (rapid change) in temperature and potentially dry streams. Due to the failure of the device, there were also missing values in both air and water temperatures. Therefore, another problem of interest is how to use statistical models to infill the missing values and remedy the oddities. Missing data imputation can be useful for the model fitting and inference. It is also helpful in extending the clustering analysis to all streams.

1.2 Literature Review

Many researchers have contributed to the modeling of bivariate spatial-temporal data. These models, from stochastic processes (Cluis, 1972) to the modern wavelet neural networks (Wang *et al.*, 2013), have advantages on different aspects. Some of them focus on the spatial or temporal correlation of water temperature. Others may be good at modeling the water and air relationship. Let $W_{s,t}$ be the maximum daily water temperature and $A_{s,t}$ be the maximum daily air temperature for site s at time t, $t = 1, 2, \ldots, T$, $s = 1, 2, \ldots, S$, where T is the number of time points and S is the number of sites in the data set. Some commonly used models are summarized in this section.

1.2.1 Models for the Water-Air Relationship

Parametric Regression Models

The linear regression model proposed by Neumann *et al.* (2003) to study the water-air relationship has the following form:

$$W = \theta_0 + \theta_1 A + \epsilon, \tag{1.1}$$

where W and A denote water temperate and air temperature and ϵ is the random error. The linear regression model can be easily implemented and all the inference tools are established and simple. Another advantage of parametric regression models is the interpretability: the parameter θ_1 in the model represents the sensitivity of the water temperature to the change of the air temperature.

A characteristic of the water-air temperature relationship in streams that affects model choice is that water temperature often remains relatively constant when the air temperature is below $0°C$. To address this issue, a nonlinear logistic regression model (Mohseni *et al.*, 1998) can be used to describe the relationship between air and water temperatures.

$$W = \mu + \frac{\alpha - \mu}{1 + e^{\gamma(\beta - A)}} + \epsilon. \tag{1.2}$$

Here α, β, γ and μ are parameters. The nonlinear relationship is appropriate because water temperature is less sensitive to air temperature in the cold seasons (air temperature less than $0°C$) due to flow and the potential for freezing. In the warm seasons (air temperature greater than $20°C$), as air temperature increases, the increase in water temperature may be small due to the high rate of evaporative cooling.

Nonparametric Models

In the recent years, the k-nearest neighbor (KNN) method has been applied to study the stream temperature and effectively improves the prediction accuracy (Benyahya *et al.*, 2008; St-Hilaire *et al.*, 2012). The KNN method consists of identifying neighbors (in both space and time) in historical data. Neighbors are defined as days or sites in the historical time series with characteristics similar to that of the day or the site. The air temperatures, called attributes, are potential predictors of water temperature. The nearest neighbors are

selected based on the similarity between attributes, which means that the nearest neighbors are the ones that have attributes with values closest to those of the simulated day. Water temperature measured on each of the k nearest neighbors are extracted from the database and a weighted average is then calculated:

$$W = \sum_{i=1}^{k} a_i W_i, \tag{1.3}$$

where the a_i are the weights with constraint $\sum_{i=1}^{k} a_i = 1$ and the W_i are the neighboring water temperature values retrieved from the database.

Neural Networks

The artificial neural network (ANN) is an advanced model to study the water-air relationship in terms of prediction (Chenard and Caissie, 2008). The ANN structure is like a black box that consists of input variables and output variables and the model inside of the black box is called the hidden layer. (See Hastie *et al.* (2009) for more details on ANN.) To model the air and water relationship, the ANN uses air temperature and temporal and spatial information as inputs and water temperature as output. It can be fitted by training (or complete) data and often provides accurate prediction and missing values imputation. Various neural networks for missing data imputation are reviewed by Coulibaly and Evora (2007).

1.2.2 Models for Temporal Data

Time series models and other stochastic models are designed to focus on the stochastic components of the data as well as the deterministic component (Caissie *et al.*, 1998; Cluis, 1972; Kothandaraman, 1971, 1972). Most of the stochastic models use a daily time step and

model variation in relationships over time. Time series models might be useful for forecasting future observations and might result in accurate predictions with rich data (Cho and Lee, 2012).

Most of the stochastic modeling of stream water temperatures, T_w, consists of separating the water temperatures into two different components: the long-term periodic or annual component T_A and the short-term component or residuals R_w (Ahmadi-Nedushan *et al.*, 2007).

$$T_w(t) = T_A(t) + R_w(t). \tag{1.4}$$

Previous studies usually use a single invariant sine function for the annual component T_A (Cluis, 1972),

$$T_A(t) = a + b\sin(\frac{2\pi}{365}(t + t_0)), \tag{1.5}$$

where a, b and t_0 are coefficients. Concomitant and lagged air temperature residuals are used as independent variables for the short-term component or residuals R_w (Kothandaraman, 1971):

$$R_w(t) = \beta_1 R_A(t) + \beta_2 R_A(t-1) + \beta_3 R_A(t-2), \tag{1.6}$$

where β_1, β_2 and β_3 are regression coefficients. $R_A(t)$, $R_A(t-1)$ and $R_A(t-2)$ are air temperature residuals at times t, $t-1$ and $t-2$.

1.2.3 Models for Spatial Data

Spatio-Temporal models (Cressie and Wikle, 2011) are designed to model the spatial and temporal correlation of multivariate time series. Most of the Spatio-Temporal models assume the variables have a joint distribution and might not be appropriate for studying the water-air relationship which requires a conditional distribution of water on air. The Spatio-Temporal

process (Cressie and Wikle, 2011) is expressed as:

$$\mathbf{X}(s,t) = \boldsymbol{\mu}(s,t) + \boldsymbol{\epsilon}(s,t). \tag{1.7}$$

Here $\mathbf{X}(s,t) = (W_{s,t}, A_{s,t})$ is the joint process of water and air temperature, $\boldsymbol{\mu}(s,t)$ is a mean process and $\boldsymbol{\epsilon}(s,t)$ is a zero mean process. Covariogram and variogram (Cressie, 1993) are often used in estimating the spatial correlation in model (1.7). For stationary processes, the spatial-temporal model can be built around kernel methods (Ver Hoef and Barry, 1998)

$$\boldsymbol{\epsilon}(s,t) = \sum_{l=1}^{L}\sum_{m=1}^{M} K(|s-w_l|, |t-t_m|)v(s,t), \tag{1.8}$$

where $K(.,.)$ is a kernel function, $v(.,.)$ is a function of s and t, and L and M are the number of space and time observations.

One of the Spatio-Temporal model, the Dynamic Linear Model (DLM) (Velasco-Cruz *et al.*, 2012) is developed based on conditional distribution and could be used to model the relationship between water and air tempreature. It has two equations, the observation equation and the system equation as follows:

$$\mathbf{W}_t = \mathbf{A}_t\boldsymbol{\theta}_t + \boldsymbol{\epsilon}_t, \tag{1.9}$$

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \boldsymbol{\eta}_t. \tag{1.10}$$

Here $\mathbf{W}_t$ and $\mathbf{A}_t$ are vectors of water and air temperature at time t, $\boldsymbol{\theta}_t$ is the dynamic coefficient at time t and $\boldsymbol{\epsilon}_t$ and $\boldsymbol{\eta}_t$ are random errors. The water-air relationship can be estimated by the mean of the posterior distribution of $\boldsymbol{\theta}_t$.

Another commonly used spatial model is the Gaussian Process model (also called Kriging)

(Cressie, 1993). It assumes that

$$W_{s,t} \sim GP(A_{s,t}\beta, \sigma^2 R), \tag{1.11}$$

where $GP(.)$ is a Gaussian Process. The mean part $A_{s,t}\beta$ in equation (1.11) is often modeled by simple linear model and the variance part $R = R_t \otimes R_s$. Here $\otimes$ is the Kronecker product and R_s and R_t are the spatial and temporal correlation of water temperature.

The spatial-temporal varying coefficient model (STVCM) (Lu *et al.*, 2009; Serban, 2011) is a relatively new approach to deal with spatial-temporal data. It has the following expression,

$$W_{s,t} = \theta_0(s,t) + A_{s,t}\theta_1(s,t) + \epsilon_{s,t}. \tag{1.12}$$

From equation (1.12), the STVCM uses a linear form between a dependent and independent variable, thus will be appropriate to describe the relationship between water and air temperature. Moreover, the nonparametric form of the coefficients in the STVCM treats the space and time covariates as inputs thus could model spatial and temporal correlation properly (Lu *et al.*, 2009; Serban, 2011). Another advantage of the VCM is that the fitting of the model is straight-forward and computationally fast (Fan and Zhang, 2008; Li *et al.*, 2014).

1.2.4 Temporal Data Smoothing

Water and air temperatures measured across time and space are usually noisy and often treated as functional data. Smoothing the water-air data by functional data analysis (FDA) methods could be beneficial in clustering streams. In this subsection, I review two commonly used methods, functional principal components and basis functions.

Functional Principal Component Analysis (FPCA)

In FPCA (Ramsay and Silverman, 2005), the weight ξ is a function of t and the ith principal component is

$$f_i = \int \xi_i(t)X(t)dt, \tag{1.13}$$

where $X(t)$ is the functional data observed at time t. Define the covariance operator as

$$v(s,t) = \text{cor}(X(s), X(t)), \tag{1.14}$$

the weight function $\xi(t)$ can be found by

$$\int v(s,t)\xi(t)dt = \lambda\xi(s), \tag{1.15}$$

where λ is an appropriate eigenvalue. The integral transform is called the covariance operator V. The eigen-equation can be expressed as

$$V\xi = \lambda\xi. \tag{1.16}$$

Berrendero *et al.* (2011) propose a between curves approach which treats multiple curves as multivariate random variables. Suppose $\Sigma(t)$ is the covariance matrix at time t, the eigenfunction $\xi(t)$can be calculated by maximizing

$$\int \xi'(t)\Sigma(t)\xi(t)dt. \tag{1.17}$$

The principal components can be obtained by

$$f_i = \xi'(t)\mathbf{X}(t). \tag{1.18}$$

Jacques and Preda (2014) adopted the method used in univariate FPCA and define the covariance operator V on a p-dimensional L_2 continuous functional space. Suppose $\{\lambda_i\}_{i\geq 1}$ is a countable set of eigenvalues of operator V, the weight function $\boldsymbol{\xi}_i(t) = (\xi_{i,1}(t), \ldots, \xi_{i,p}(t))$ is the eigenfunction corresponding to λ_i.

$$V\boldsymbol{\xi}_i(t) = \lambda_i \boldsymbol{\xi}_i(t). \tag{1.19}$$

The principal components $f_{i,j}$ are defined as the projections of $\mathbf{X}$ on the eigenfunctions.

$$f_i = \int_0^T < \mathbf{X}(t), \boldsymbol{\xi}_i(t) >_{\mathbb{R}^p} dt = \int_0^T \sum_{j=1}^{p} X_j(t)\xi_{i,j}(t)dt. \tag{1.20}$$

Here additivity is assumed; a more general setting should be:

$$f_{i,j} = \int_0^T < X(t), \xi_{i,j}(t) >_{\mathbb{R}} dt = \int_0^T X_j(t)\xi_{i,j}(t)dt \tag{1.21}$$

Basis expansion

Basis expansion is commonly used in mathematical analysis; examples include the Taylor expansion and Fourier expansion. In statistical analysis, the expression of functional data is usually not clear so it is better to study the data structure and extract the basis function from the data. Chiou and Li (2007) propose using a Karhunen-Loève expansion for random functions,

$$X(t) = \mu(t) + \sum_{i=1}^{\infty} \xi_i(X)\rho_i(t). \tag{1.22}$$

It is the same approach as the principal component analysis, where $\rho_i(t)$ is the eigenfunction of the covariance operator and $\xi_i(X)$ is the principal component score. Ray and Mallick

(2006) propose a wavelet basis expansion for functional data clustering

$$f_i(t) \approx \beta_{i00}\phi_{00}(t) + \sum_{j=1}^{J}\sum k = 1^{2^j-1}\beta_{ijk}\psi_{jk}(t), \tag{1.23}$$

where (ϕ, ψ) is the wavelet basis and β is the coefficient. Abraham *et al.* (2003) fit a B-spline to the functional data

$$X(t) = \sum_{l=1}^{K+d+1} \beta_l B_l(t). \tag{1.24}$$

Kayano *et al.* (2010) use orthonormalized Gaussian basis

$$u(t) = \sum_{m=1}^{M} c_m \phi_m(t), \tag{1.25}$$

where $\phi_m(t) = \exp(-\frac{(t-k_{m+2})^2}{2\gamma^2})$.

1.3 Outline of this Dissertation

Two essential problems in studying the water-air relationship are interpreting this relationship and using air temperature to predict water temperature. In chapter 2, I solve both problems for individual streams. I propose the use of a varying coefficient model (VCM) to study the relationship between daily maximum water temperature and daily maximum air temperature (Fan and Zhang, 2008). The VCM can provide a meaningful interpretation and often results in a fitted model with accurate prediction. In the later chapters, I consider the spatial correlation of the streams and study them simultaneously. In chapter 3, I focus on the first problem and use the VCM to obtain more meaningful results. Specifically, I apply the VCM to multiple streams, cluster streams based on the water-air relationship and explore how climate and landscape variables are related to the water-air relationship. In chapter

4, I focus on the missing data problem and develop a spatial-temporal VCM (STVCM) for multiple streams. This model is used to fill the missing values in the water temperature. For some extreme cases with big gaps of missing values, the single VCM in chapter 2 does not perform very well for missing data imputation. The STVCM is more general and often can result in accurate imputation. The use of the STVCM for missing values can extend our study to all the streams under monitoring.

In chapter 2, I apply the VCM to daily maximum temperature data from 10 native brook trout streams in southeast United States and show that the VCM outperforms the logistic or linear models in prediction. The VCM is an effective tool for exploring the dynamic feature of the data. The key idea of the VCM is to use a parametric model but with varying coefficients. Therefore, the VCM can also be used to interpret air and water temperature relationships over time.

In chapter 3, I cluster 62 streams based on water-air relationship by using the intercept and slope curves from the VCM. I propose a weighted distance measure for clustering based on the water-air temperature relationship. Specifically, I cluster streams based on the bivariate curves that summarize the relationship with spatial restrictions. I use the VCM to quantify the water-air temperature relationship for individual streams and develop a weighted Canberra distance to measure the distance between the intercept and slope curves. Then I adopt the idea in Giraldo *et al.* (2012) to adjust the weighted distance by a scale of spatial distance using the variogram (Cressie, 1993). This model-based spatially-adjusted distance makes clusters both statistically interpretable and spatially meaningful.

In chapter 4, I propose a spatial-temporal VCM (STVCM) to incorporate the spatial correlation of the data. The STVCM takes both spatial and temporal variation of the water temperature into account. I propose a novel estimation method that emphasizes the time effect and treats the space effect as a varying coefficient for the time effect. The STVCM

is effective in handling large gaps of missing values and can obtain accurate imputation results. The STVCM are used to infill all the missing values in the 156 streams. I illustrate the merits of the proposed STVCM on missing data imputation by applying the clustering algorithm used in chapter 3 to all 156 streams.

Chapter 2

A Varying Coefficient Model for Single Stream.

Li H, Deng X, Kim D-Y, Smith E. P, 2014. Modeling maximum daily temperature using a varying coefficient regression model. Water Resources Research 50: 3073-3087.

2.1 Introduction

Statistical models have been used to study the air and water temperature relationship (Benyahya *et al.*, 2007) and tend to work well for weekly mean temperature. Among different statistical methods, parametric regression models such as the linear regression model are easily implemented; all the inference tools are established and simple. Another advan-

tage of parametric regression models is the interpretability. Parameters in the model have meaningful interpretation. Neumann *et al.* (2003), for example, developed a linear regression method to model daily maximum stream temperature in terms of maximum air temperature and other related predictors for the Truckee River in California and Nevada. The majority of the statistical models focus on weekly or greater time steps. This is partly a result of the data collection process and validity of the model. Caissie (2006) pointed out that linear regression models were more valid at weekly or monthly scales than at a daily scale. At these scales autocorrelation tends to be less than that at a daily scale and the normality assumption is reasonable when average temperature is used. A characteristic of the water-air temperature relationship in streams that affects model choice is that water temperature often remains relatively constant when the air temperature is below $0°C$. In this case, the simple parametric linear regression model might not be a good model, as the relationship becomes nonlinear. To address this weakness, nonlinear parametric models might be more applicable.

The most commonly used nonlinear parametric regression model for the air-water temperature relationship is the logistic model (Mohseni *et al.*, 1998). The model describes the relationship between air and water temperatures as

$$W = \mu + \frac{\alpha - \mu}{1 + e^{\gamma(\beta - A)}}, \tag{2.1}$$

where W is the measured stream temperature, A is the measured air temperature and α, β, γ and μ are parameters. Compared to the linear regression model, the logistic model provides a good explanation of the flat patterns in water temperature for low ($< 0°C$) and sometimes for high ($> 20°C$) air temperature. The nonlinear relationship is appropriate because water temperature is less sensitive to air temperature in the cold seasons (air temperature less than $0°C$) due to flow and the potential for freezing. In the warm seasons (air temperature

greater than $20°C$), as air temperature increases, the increase in water temperature may be small due to the high rate of evaporative cooling. As the S-shaped logistic function flattens at both ends of the range of air temperature, the logistic model easily accommodates these phenomena. A physical interpretation was also given in Mohseni and Stefan (1999). Several recent studies have used the logistic model for modeling streams (Webb *et al.*, 2003; Mayer, 2012). There are two potential problems that limit the use of the logistic model, especially, at the daily time scale. First, in some situations, daily water and air temperatures have high variability at certain temperatures, and this results in poor prediction (despite reasonably high goodness-of-fit statistics). If this variability is associated with seasonality, the standard logistic model would not address this, resulting in greater variance and weaker predictive ability. This is described as hysteresis (stream temperature being different for the same air temperature at different times of the year) in Mohseni *et al.* (1998). Second, daily water temperatures are likely to have high autocorrelation over time. The autocorrelation affects inference and confidence in parameter estimates. This dependence across time is not taken into account by the standard logistic model, hence the variance estimates, prediction and hypothesis testing might not be accurate at a daily time scale.

Compared to the parametric model, nonparametric regression models often provide good prediction of water temperature. The nonparametric models usually have simple assumptions and therefore are widely applicable in different situations. Modern approaches to handle temporal correlation are possible and the nonparametric models can effectively improve the prediction accuracy (Chenard and Caissie, 2008). An example of a nonparametric model is the k-nearest neighbor method which was used to predict and forecast water temperature (Benyahya *et al.*, 2008; St-Hilaire *et al.*, 2012). One drawback of such nonparametric models is the difficulty in interpreting the parameters in the model.

Time series models and models based on stochastic processes were designed to focus on the

stochastic components of the data as well as the deterministic component (Caissie *et al.*, 1998; Cluis, 1972; Kothandaraman, 1971, 1972; Stefan and Preud'homme, 1993). Most of the stochastic models use a daily time step and model variation in relationships over time. Time series models might be useful for forecasting future observations and with rich data might result in accurate predictions (Cho and Lee, 2012). One drawback of time series models is that they might not directly model the relationship between air temperature and water temperature and hence parameters in the model would not be helpful in determining the strength of the relationship and the sensitivity of water temperature to air temperature. For example, Ahmadi-Nedushan *et al.* (2007) added lagged air temperature residuals to their model of water temperature rather than actual air temperature. Cho and Lee (2012) linked air and water temperature by assuming that the ratio and/or difference of the harmonic coefficients between the air and water temperatures remained constant.

This work is part of a larger study whose goal is to identify streams in southeastern United States that are likely to lose trout due to increased temperature. Hence measuring and visualizing sensitivity of water temperature to changes in air temperature is an important goal. In this section, I propose the use of a time varying coefficient model (VCM) to study the relationship between daily water temperature and daily air temperature (see Fan and Zhang (2008) for an overview of the VCM). The VCM is an effective tool for exploring the dynamic feature of the data and has been widely applied in different areas such as ecology (Ferguson *et al.*, 2007, 2009) and medicine (Cheng *et al.*, 2009). The key idea of the VCM is to use a parametric model but with time varying coefficients. The parametric property of the VCM provides meaningful interpretation of the sensitivity of stream water temperature to changes in air temperature, aids in understanding how sensitivity varies over time and provides a way to compare sensitivity across streams.

The varying coefficient model is a useful model in situations where there is a significant

amount of variability in water temperature that is not accounted for by air temperature but might be related to unmeasured factors that vary over time. For example, the natural cycle of temperature suggests that the water temperature in streams will be different for the same value of air temperature at different times of the year. Thus an air temperature of $10°C$ might be associated with a different water temperature in the spring than in the fall. In this situation, both the linear regression and the nonlinear logistic regression might result in reasonable model fit statistics but may not provide accurate prediction of water temperature based on air temperature as these models overlook the time (seasonal) information in the data. For similar air temperatures at different times of the year, daily water temperature varies considerably but the resulting predictions of water temperature would be almost constant. The proposed VCM surmounts this difficulty by taking the time information into account and modeling the temporally dynamic pattern of the air-water temperature relationship through varying coefficients. Therefore, it can achieve more accurate predictions when the same air temperature occurs at different time of the year. Moreover, the proposed modeling strategy better explains the sensitivity of water temperature to the air temperature across time, providing a more comprehensive understanding of the air-water relationship in different time periods. Modeling the relationship with additional terms other than time varying coefficients (e.g., seasonal terms) is possible but might be complicated, especially when using the logistic model, and requires choices about the number of seasonality terms to add to the model and the period associated with seasonality.

In this section, I apply the VCM to daily maximum temperature data from 10 native brook trout streams in southeast United States and show that the VCM has predictions that are superior to the logistic or linear models. In addition, the VCM can be used to interpret the air and water temperature relationship over time.

Chapter 2. A Varying Coefficient Model for Single Stream.: *Li H, Deng X, Kim D-Y, Smith E. P, 2014. Modeling maximum daily temperature using a varying coefficient regression model. Water Resources Research 50: 3073-3087.*

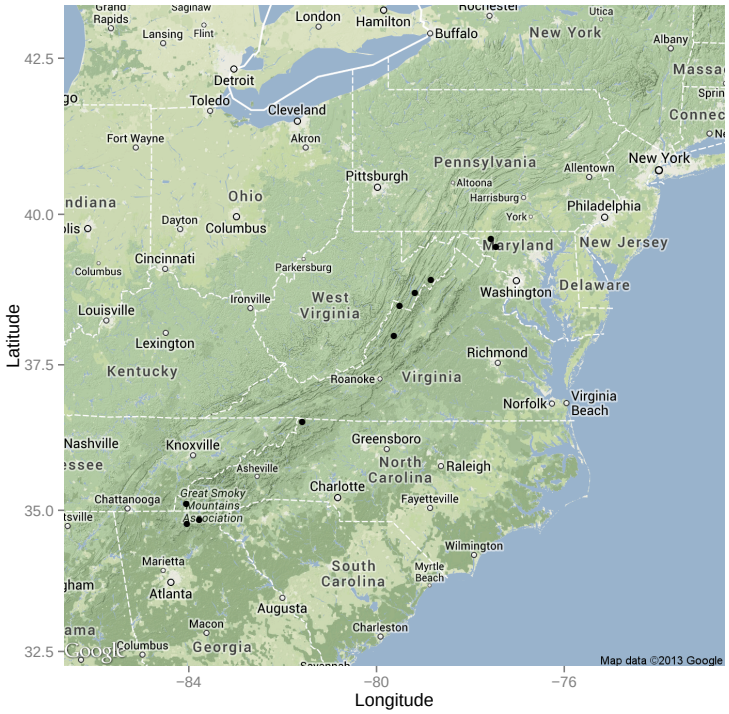

Figure 2.1: Location of 10 streams (black dots are the locations). From north to south, the site ID's are: Site 5, Site 6, Site 10, Site 9, Site 2, Site 1, Site 8, Site 7, Site 4 and Site 3.

2.2 Study Area and Data

In this chapter, 10 streams located in the states of Maryland, West Virginia, Virginia, North Carolina and Georgia were used, and their locations are shown in Figure 2.1. For each site, I extracted one complete year of data with the same starting date (14 December 2010) and ending date (14 December 2011). To summarize the daily data, daily maximum values were used. The data that were input into the models consisted of 366 paired daily maximum air and water temperatures for each site.

Chapter 2. A Varying Coefficient Model for Single Stream.: *Li H, Deng X, Kim D-Y, Smith E. P, 2014. Modeling maximum daily temperature using a varying coefficient regression model. Water Resources Research 50: 3073-3087.*

2.3 Varying Coefficient Model

Let W_t be the maximum water temperature and Z_t be the maximum air temperature in day t, $t = 1, 2, \ldots, T$ and $T = 366$ is the total number of days in the data set. Without loss of the generality, I use centered air temperature, i.e., $A_t = Z_t - \bar{Z}$, where $\bar{Z} = \frac{1}{T}\sum_{t=1}^{T} Z_t$. I consider the following varying coefficient model for the air-water temperature relationship as

$$W_t = \theta_0(t) + A_t\theta_1(t) + \epsilon_t, \tag{2.2}$$

where $\theta_0(t)$ and $\theta_1(t)$ are varying intercept and slope coefficients and ϵ_t is the error term in the model. Exploratory analysis using lag-1 autocorrelation of residuals and the Shapiro-Wilk test (Kutner *et al.*, 2004) indicated that the error terms were neither normally distributed nor independent. Levene's test (Kutner *et al.*, 2004) indicated that the assumption of constant variance in the error terms was reasonable. Based on these analyses, I assume $E(\epsilon_t) = 0$ and $\text{var}(\epsilon_t) = \sigma^2$. Note that the proposed model has a similar format as the linear regression model $W_t = \theta_0 + A_t\theta_1 + \epsilon_t$. These varying coefficients can be interpreted as the dynamic feature of the air-water temperature relationship. That is, at different times and seasons, water temperature and sensitivity of water temperature to changes in air temperature can be different. This allows for a seasonal effect on the air-water temperature relationship. The full VCM is thus useful for following changes in the maximum water temperature over time (through the intercept) and for measuring the local sensitivity of the relationship (through the slope). Variants of the varying coefficient model can accommodate flexibility in model interpretation. For instance, one can consider a semi-varying coefficient model, which has the following form

$$W_t = \theta_0 + A_t\theta_1(t) + \epsilon_t. \tag{2.3}$$

In model (2.3), the intercept θ_0 is a fixed parameter. Hereafter, I call model (2.2) the full VCM and model (2.3) the semi-VCM. The fixed intercept θ_0 in model (2.3) represents the average maximum water temperature for the whole period. The varying slope $\theta_1(t)$ in model (2.3) represents variation in slope relative to the simple linear regression. The time varying slope provides information about the degree of deviation from a common slope. Also the times when the variation is the greatest may be relevant. Therefore, depending on the objective, one can choose either the full VCM or the semi-VCM to describe the local or global behavior of the water temperature, given the air temperature. For presentation convenience, the full VCM will be used to illustrate the details of the proposed methodology; the equations can be easily adapted to the semi-VCM.

2.3.1 Estimation of Coefficients

Two popular methods for estimating the varying coefficients $\theta_0(t)$ and $\theta_1(t)$ are kernel-local polynomial smoothing and smoothing spline methods (Fan and Zhang, 2008). In this work, I adopt the penalized spline regression method as it has parsimonious parameter expression with easy interpretation (Ruppert *et al.*, 2003). Specifically, the penalized spline approach assumes the varying coefficients have the form

$$\theta_0(t) = \sum_{i=1}^{K} \alpha_i b_i(t), \tag{2.4}$$

$$\theta_1(t) = \sum_{i=1}^{K} \beta_i b_i(t). \tag{2.5}$$

where $\{b_1(t), \cdots, b_K(t)\}$ is the set of basis functions and α_i and β_i, $i = 1, 2, \cdots, K$ are parameters. K is the number of basis functions for each varying coefficient.

The key is to estimate the coefficients α_i and β_i such that the estimates make the series of varying coefficients $\theta_0(t)$ and $\theta_1(t)$ smooth. To achieve this purpose, I use least squares with a penalty to encourage smoothness (Hoover *et al.*, 1998). Specifically, I obtain the estimates by minimizing

$$\sum_{t=1}^{T}(W_t - \theta_0(t) - A_t\theta_1(t))^2 + \lambda(\int \theta_0''(t)^2dt + \int \theta_1''(t)^2dt), \tag{2.6}$$

where $\theta_0(t)$ and $\theta_1(t)$ were given in (2.4) and (2.5) and λ is a tuning parameter for controlling the smoothness of varying coefficients. Given the observed data, I can write the vector of water temperatures as $\mathbf{W} = (W_1, \cdots, W_T)'$ and the vector of air temperature as $\mathbf{A} = (A_1, \cdots, A_T)'$. Denote $\mathbf{X} = (\mathbf{b}_1, \cdots, \mathbf{b}_K, \mathbf{b}_1 \circ \mathbf{A}, \cdots, \mathbf{b}_K \circ \mathbf{A})_{T\times(2K)}$, where $\mathbf{b}_i = (b_i(1), \cdots, b_i(T))', i = 1, \cdots, T$, and $\circ$ is the Schur product. It is easy to obtain the estimated parameter set, given λ, as

$$(\hat{\alpha}_1, \cdots, \hat{\alpha}_K, \hat{\beta}_1, \cdots, \hat{\beta}_K)' = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{D})^{-1}\mathbf{X}'\mathbf{W}, \tag{2.7}$$

where $\mathbf{D}$ is a $2K$ by $2K$ diagonal penalty matrix with diagonal elements either 0 or 1. Consequently, the estimates of the varying coefficients are $\hat{\theta}_0(t) = \sum_{i=1}^{K}\hat{\alpha}_i b_i(t)$ and $\hat{\theta}_1(t) = \sum_{i=1}^{K}\hat{\beta}_i b_i(t)$.

2.3.2 Tuning Parameter Selection

Note that there is a tuning parameter λ that controls the smoothness of the varying coefficients and influences the model fit. To select an optimal tuning parameter λ, one commonly used approach is leave-one-out cross-validation (LOOCV) suggested by Hoover *et al.* (1998). Let $\hat{\theta}_0^{(-i)}(t)$ and $\hat{\theta}_1^{(-i)}(t)$ be the varying coefficients estimated by minimizing (2.6) by using

data with the ith observation deleted. The $LOOCV(\lambda)$ is defined as

$$LOOCV(\lambda) = \sum_{i=1}^{T}(W_i - \hat{\theta}_0^{(-i)}(i) - A_i\hat{\theta}_1^{(-i)}(i))^2. \tag{2.8}$$

One can choose the optimal tuning parameter λ_{loocv} as the value minimizing $LOOCV(\lambda)$. However, there are $T = 366$ data points for each site, and implementing such a method can be computationally expensive. To circumvent this difficulty, I adopt the generalized cross-validation method (GCV) (Wahba, 1990) to approximate the leave-one-out cross-validation method. Here the $GCV(\lambda)$ is defined as

$$GCV(\lambda) = (\hat{\mathbf{W}} - \mathbf{W})'(\hat{\mathbf{W}} - \mathbf{W})/(1 - \mathrm{tr}(\mathbf{S}_\lambda)/T), \tag{2.9}$$

where $\mathbf{S}_\lambda = \mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda\mathbf{D})^{-1}\mathbf{X}'$ is the so-called hat matrix, $\hat{W}_t = \hat{\theta}_0(t) + A_t\hat{\theta}_1(t)$ is the fitted water temperature based on the proposed model using the entire data set and $\hat{\mathbf{W}} = (\hat{W}_1, \cdots, \hat{W}_T)'$. Then the optimal tuning parameter λ_{GCV} is the one minimizing $GCV(\lambda)$. The GCV often gives a very reasonable approximation to the LOOCV and can effectively reduce the computational time (Ruppert *et al.*, 2003).

2.3.3 Model Assessment: Fitting and Inference

To evaluate the goodness-of-fit for the proposed VCMs, the Nash-Sutcliffe coefficient (NSC) (Nash and Sutcliffe, 1970) is used as a performance measure

$$\mathrm{NSC} = 1 - \frac{\sum_{t=1}^{T}(W_t - \hat{W}_t)^2}{\sum_{t=1}^{T}(W_t - \bar{W})^2}, \tag{2.10}$$

where $\bar{W} = \frac{1}{T}\sum_{t=1}^{T} W_t$. In the linear regression model, the NSC is equivalent to the coeffi-

cient of determination, i.e., R^2 (Rencher and Schaalje, 2008). To compare the fit of different models, I define the relative NSC (RelNSC) using the linear regression model as the baseline, i.e.,

$$\text{RelNSC} = \frac{\text{NSC} - \text{NSC}_0}{\text{NSC}_0} \times 100\%, \tag{2.11}$$

where NSC_0 is the NSC of the linear regression model. A higher RelNSC means a better fit. Note that the RelNSC of the linear regression model is 0%. Here I remark that for fitting the nonlinear logistic model, I followed the two-step iterative estimation method used in Mohseni *et al.* (1998): iterate the step of estimating α and μ by least squares and the step of estimating β and γ by Newton's method.

Although the VCM has a varying coefficient form, inference and testing for VCM is straightforward by using parameter estimation in (2.7). Similar to routine regression modeling one can take advantage of model nesting to evaluate model fit. That is, one can conduct a statistical hypothesis test to check whether the VCM is significant relative to the linear regression model or some other reduced model. Under such a consideration, the null hypothesis is that none of the coefficients are time-varying, i.e.,

$$H_0 : \theta_0(t) = \theta_0 \text{ and } \theta_1(t) = \theta_1 \tag{2.12}$$

versus the alternative that at least one of the coefficients is not constant. Suppose the two models under the null and the alternative hypothesis are M_0 and M_1, respectively

Chapter 2. A Varying Coefficient Model for Single Stream.: *Li H, Deng X, Kim D-Y, Smith E. P, 2014. Modeling maximum daily temperature using a varying coefficient regression model. Water Resources Research 50: 3073-3087.*

$$M_0 : W_t = \theta_0 + \theta_1 A_t + \epsilon_t, \tag{2.13}$$

$$M_1 : W_t = \theta_0(t) + \theta_1(t) A_t + \epsilon_t. \tag{2.14}$$

Because of autocorrelation and lack of normality, a standard F test is not appropriate. To test the difference between the two models, I apply a block bootstrap goodness-of-fit test based on the comparison of sum of the squared residuals for the varying coefficient model (Huang *et al.*, 2002). Note that there is no assumption of a specific distribution for ϵ_t in (2.1), hence bootstrap-based testing should be more robust and reliable. Suppose $\hat{\theta}_0$ and $\hat{\theta}_1$ are parameter estimates for the model M_0, and $\hat{\theta}_0(t)$ and $\hat{\theta}_1(t)$ are parameter estimates for the model M_1. I then can calculate the residual sum of squares for each model respectively

$$RSS_0 = \sum_{t=1}^{T} (W_t - \hat{\theta}_0 - \hat{\theta}_1 A_t)^2, \tag{2.15}$$

$$RSS_1 = \sum_{t=1}^{T} (W_t - \hat{\theta}_0(t) - \hat{\theta}_1(t) A_t)^2. \tag{2.16}$$

Using RSS_0 and RSS_1, I can define the test statistic $G = \frac{(RSS_0 - RSS_1)/(2K-2)}{RSS_1/(T-2K)}$, which is the standard F test statistic under the assumption that the error term ϵ_t's are independent and identically normally distributed (Rencher and Schaalje, 2008). A large value of the test statistic implies a significant difference between the two models. To assess the level of significance accurately, I adopt the block bootstrap method to compute the critical value (Huang *et al.*, 2002). Because water temperatures and residuals are highly autocorrelated through time, the ordinary bootstrap is not appropriate and the block bootstrap (Kunsch, 1989) is

therefore more appropriate. The proposed bootstrap testing procedure is summarized as follows:

1. Calculate $G = \frac{(RSS_0 - RSS_1)/(2K-2)}{RSS_1/(T-2K)}$ using the original data.

2. Let

$$\hat{\epsilon}_t = W_t - \hat{\theta}_0(t) - \hat{\theta}_1(t)A_t$$

be the residual under the alternative hypothesis. Split the sample $\{\hat{\epsilon}_1, \cdots, \hat{\epsilon}_T\}$ into $T-m+1$ overlapping blocks of length m: observation 1 to m will be block 1, observation 2 to $m+1$ will be block 2 and so on. Resample T/m blocks with replacement from the $T-m+1$ blocks. Aligning these T/m blocks in the order they were selected will give the new sample: $\{\hat{\epsilon}_1^*, \cdots, \hat{\epsilon}_T^*\}$. Note that the last block has extra points when T/m is not integer.

3. Let

$$W_t^* = \hat{\theta}_0 + \hat{\theta}_1 A_t + \hat{\epsilon}_t^*$$

be the pseudo-responses under the null hypothesis.

4. Repeat step 2 and step 3 B times to obtain B bootstrap samples.

5. From each bootstrap sample, calculate $G^b = \frac{(RSS_0^b - RSS_1^b)/(2K-2)}{RSS_1^b/(T-2K)}, b = 1, 2, \cdots, B$. Reject H_0 if the value of statistic G is greater than or equal to the $\{100(1-\alpha)\}$ percentile of G^b, where α is the significance level.

2.3.4 Model Assessment: Prediction

To evaluate the prediction performance of the VCM compared to the parametric models, for the data $\mathcal{X} = \{(A_t, W_t) : t = 1, \ldots, T\}$ $(T = 366)$ from each site, I randomly partition

two-thirds of the data as the training set and the remaining as the test set. For each model compared, I calculate the root mean squared errors (RMSE) of the test set using the model estimated by the training set, which is defined as

$$RMSE = \sqrt{\frac{1}{|\mathcal{Y}| - \mathrm{df}} \sum_{t \in \mathcal{Y}} (W_t - \hat{W}_t)^2}, \tag{2.17}$$

where $\mathcal{Y}$ is the test set and $|\mathcal{Y}|$ is the number of the observations in set $\mathcal{Y}$. In Equation (2.17), df is the degree of freedom in the model, where df is 14 for full VCM (there are 7 bases for each of the two varying coefficients), 8 for semi-VCM (since the intercept is fixed and there 7 bases for slope), 4 for the logistic model and 2 for the linear regression model.

2.4 Results

In this section, I fitted both the full VCM and semi-VCM to the data from 10 sites, respectively. The proposed methods were compared to the linear regression model and the nonlinear logistic model. The merits of the proposed VCM method will be examined through fit statistics, prediction, and interpretation of the relationship between water and air temperature. In this work, all the analyses were implemented by R software (version 2.15.3) (Hornik, 2013).

Figure 2 illustrates the data from Site 5. In Figure 2.2, I can see that although water and air temperature follow a similar pattern, the variation of water and air temperatures is large in spring and fall. Therefore, the prediction of the water temperature may be inaccurate if one ignores the temporal information in the model.

Chapter 2. A Varying Coefficient Model for Single Stream.: *Li H, Deng X, Kim D-Y, Smith E. P, 2014. Modeling maximum daily temperature using a varying coefficient regression model. Water Resources Research 50: 3073-3087.*

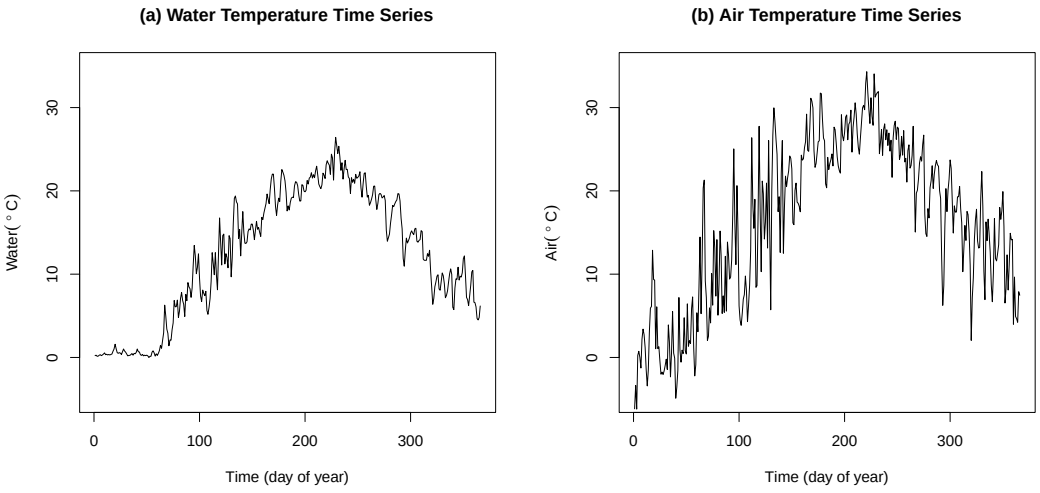

Figure 2.2: Plots of water and air temperatures for Site 5.

2.4.1 Basis Selection and Fitting

For fitting the data using the proposed VCM method, I need to choose the form of basis functions in (2.4) and (2.5). I considered three different polynomial bases as the candidates: linear, quadratic and cubic splines (Ruppert *et al.*, 2003). Analysis of the data using these three bases resulted in similar curves for the estimated varying coefficients and also similar prediction performance. In this work, I chose to use quadratic splines because they resulted in smoother coefficient curves than linear splines, and had fewer parameters compared with cubic splines. Specifically, the bases I used are

$$\{1, t, t^2, (t-\xi_1)_+^2, \cdots, (t-\xi_N)_+^2\},$$

where, $\xi_1, \xi_2, \cdots, \xi_N$ are N knots and $(t-\xi_n)_+, n = 1, 2, \cdots, N$ are the splines with $(t-\xi_n)_+ = t - \xi_n$ if $t \geq \xi_k$ and $(t-\xi_n)_+ = 0$ if $t < \xi_n$.

To select the number of knots, I fixed $N = 4$ which corresponds to four seasons. The location of the knots are evenly distributed in time, i.e., $(\xi_1, \xi_2, \xi_3, \xi_4) = (74, 147, 220, 293)$.

Here numbers represent the day of the year, staring from 14th December 2010, i.e., 14th December 2010 = 1. I also estimated the proposed model and conducted prediction by using $N = 12$. It provided similar results to those using $N = 4$. Note that choosing a relatively small number of knots reduces the number of parameters in the model and reduces the curvature of the coefficient curves.

The first three columns of Table 2.1 show the RelNSC values for the full VCM, the semi-VCM and the nonlinear logistic model for the 10 sites. Comparing the results with the nonlinear logistic model, both full VCM and semi-VCM have larger RelNSC values since they dynamically quantify the relationship between water and air temperatures. Moreover, the full VCM fits the data better than the semi-VCM. Note that the use of a varying intercept could further adjust the mean water temperature at each time point. Therefore, the full VCM better captures the variance of water temperature leading to higher RelNSC values.

As the proposed VCM method gives high RelNSC values, it is also important to check whether the VCM is over-fitting. I applied the goodness-of-fit test procedure described in section 2.3.3 to the full VCM and semi-VCM for the 10 sites. The testing results show that both the full VCM and the semi-VCM are significantly different from the linear regression model (see the last two columns of Table 2.1 for p-values) and the full VCM is significantly different from the semi-VCM (with p-value less than 0.001 for all ten sites) at the level $\alpha = 0.05$ for each site. Such a finding implies that the time information in the VCM is important in explaining the variation in the data.

2.4.2 Prediction

Both full VCM and semi-VCM have more accurate predictions compared to the linear regression model and the nonlinear logistic model. In section 2.3.4, I introduced a procedure

Table 2.1: Model assessment results: fit statistics and hypothesis tests. The first three columns show the values of RelNSC fitted by using the three models (nonlinear logistic model, semi-VCM and full VCM). The last two columns show p-values from hypothesis testing of semi-VCM and full VCM versus a linear regression model.

site	logistic	semi-VCM	full VCM	semi-VCM p-value	full VCM p-value
5	1%	5%	14%	<0.001	<0.001
6	2%	7%	15%	<0.001	<0.001
10	0%	3%	7%	<0.001	<0.001
9	0%	7%	13%	<0.001	<0.001
2	0%	2%	11%	<0.001	<0.001
1	0%	3%	9%	<0.001	<0.001
8	2%	7%	15%	<0.001	<0.001
7	1%	6%	14%	<0.001	<0.001
4	0%	5%	26%	<0.001	<0.001
3	2%	10%	21%	<0.001	<0.001

for prediction validation. I repeated this procedure 250 times for each site and reported the mean and the standard deviation of predicted RMSE for the 10 sites in Table 2.2. Figure 2.3 shows boxplots for RMSEs for different models based on 250 data partition for Site 1. (Boxplots for other sites could be found in supporting materials.) From the results in Table 2.2 and Figure 2.3, the values of RMSE for the VCM methods are smaller than those for the linear regression model and the nonlinear logistic model. The full VCM also has more accurate predictions than the semi-VCM. The lower RMSEs indicate that the prediction accuracy of the VCM is improved through the ability to incorporate the seasonal trend of the water temperature. An interesting observation is that, for Site 9, the prediction of the linear regression model is slightly better than the nonlinear logistic model. An explanation is that the air temperature is not below zero for a long enough period so the logistic model does not describe the relationship as well as a linear regression.

The inaccurate predictions for the linear regression model and the logistic model could be

partly explained by the large variability in water temperature with respect to fixed values of air temperature. For elaboration, I extracted 20 data points with air temperature around $20°C$ from the data for Site 3. Figure 2.4(a) shows the corresponding water temperatures. Clearly there is a large variation in the water temperature with values varying from $7°C$ to $20°C$ (the circles in Figure 2.4(a)). As shown in Figure 2.4(c) and Figure 2.4(d), both the linear regression model and the logistic model predict the water temperature to be around $13°C$. The two models ignore the time information in the data. They predict the water temperature by only using air temperature information and hence predict an almost constant value for water temperature. Taking the time information into account, the full VCM gives much more accurate predictions even at a fixed air temperature level. Figure 2.4(b) shows that the predictions given by the full VCM are very close to the true values of water temperature. Thus, although the NSC is high for the linear regression model and the logistic model, the prediction error could be large due to the ignoring of temporal information. The maximum prediction error across the ten sites for the linear regression model and the logistic model is around $7°C$. For the full VCM, it is reduced to $3°C$.

2.4.3 Model Interpretation

The proposed VCM method not only provides accurate predictions, but also gives meaningful interpretations. In the full VCM, the varying intercept term, $\theta_0(t)$, represents the mean water temperature at time t, and the varying slope term, $\theta_1(t)$, represents the local sensitivity of water temperature to changes in air temperature at time t. To elucidate, I reanalyzed the data of Site 5 for further illustration. First, I divided the data set into 12 disjoint subsets; each subset consists of about 30 data points from 30 consecutive days in a month or so. Then I analyzed these 12 monthly data sets using the linear regression model. Figure 2.5 shows the estimated intercept and slope for the 12 models for Site 5. The 12 intercepts and

Chapter 2. A Varying Coefficient Model for Single Stream.: *Li H, Deng X, Kim D-Y, Smith E. P, 2014. Modeling maximum daily temperature using a varying coefficient regression model. Water Resources Research 50: 3073-3087.*

Table 2.2: Predicted RMSE for 10 sites for VCM, logistic model and linear regression model. The mean and (standard deviation) are based on 250 different test sets for each site.

Site	Linear	Logistic	semi-VCM	full VCM
1	3.24 (0.19)	3.16 (0.18)	2.83 (0.17)	1.49 (0.10)
2	2.57 (0.13)	2.43 (0.13)	2.05 (0.11)	1.01 (0.06)
3	1.66 (0.08)	1.66 (0.09)	1.44 (0.09)	1.08 (0.08)
4	1.92 (0.09)	1.90 (0.10)	1.50 (0.09)	1.09 (0.08)
5	2.68 (0.19)	2.64 (0.18)	2.46 (0.18)	1.23 (0.07)
6	2.55 (0.14)	2.52 (0.14)	2.14 (0.15)	1.04 (0.06)
7	2.85 (0.16)	2.66 (0.15)	2.19 (0.14)	0.94 (0.06)
8	2.55 (0.13)	2.44 (0.13)	2.07 (0.12)	1.11 (0.06)
9	2.71 (0.16)	2.73 (0.16)	2.44 (0.16)	0.94 (0.05)
10	3.52 (0.18)	3.37 (0.19)	2.77 (0.18)	1.23 (0.07)

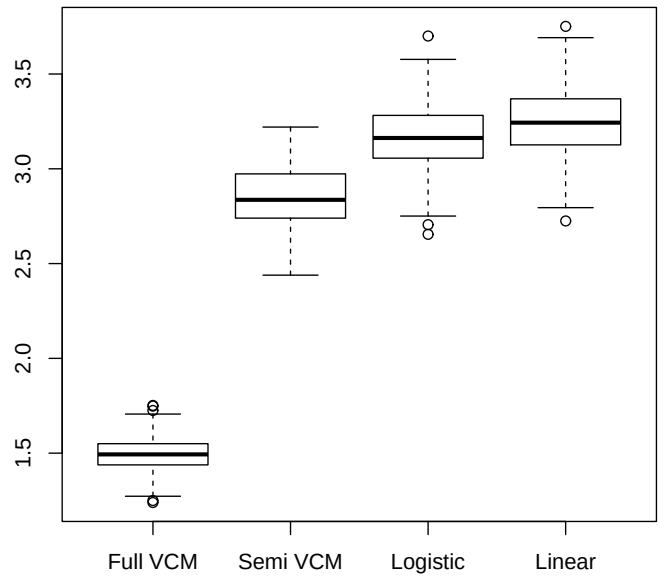

Figure 2.3: Boxplots of RMSEs for 250 random training and testing samples for four models for Site 1.

Chapter 2. A Varying Coefficient Model for Single Stream.: *Li H, Deng X, Kim D-Y, Smith E. P, 2014. Modeling maximum daily temperature using a varying coefficient regression model. Water Resources Research 50: 3073-3087.*

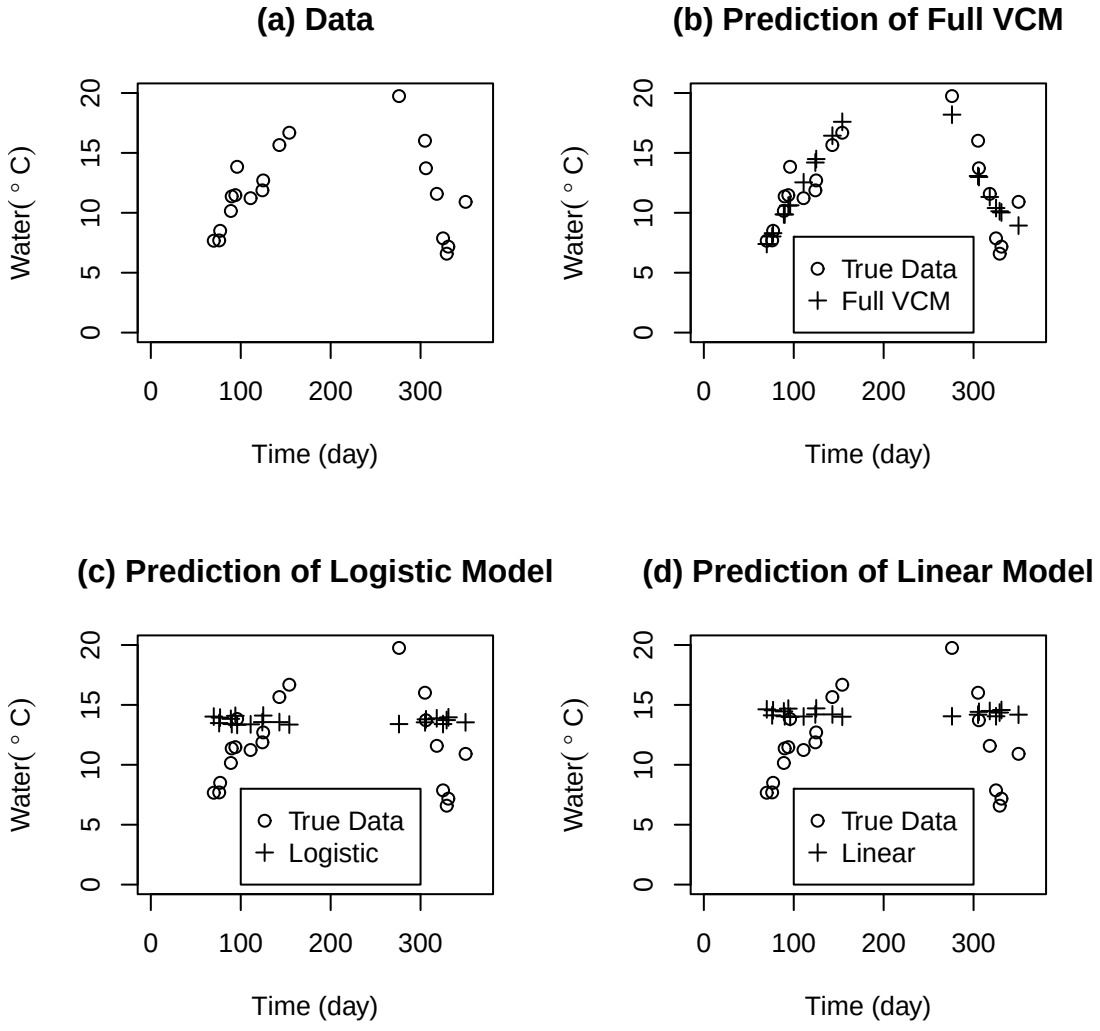

Figure 2.4: Prediction comparison of full VCM with linear regression model and logistic model for selected points where the air temperature is fixed around $20°C$.

slopes were plotted at the last time point for each of the 12 subsets and were connected by straight lines to have a piecewise form. In addition, Figure 2.5 shows the coefficients for the full VCM, semi-VCM and the linear regression model. For the varying intercept and the varying slope in the full VCM, I can see that the coefficients are very close to those in the 12 month piecewise linear regression. It shows that the full VCM can describe the local dynamic relationship between water temperature and air temperature. Therefore, the full VCM is particularly useful if one wants to study or predict water temperature in any particular season or time period over the year.

By incorporating the time information into the model, the full VCM is also able to automatically investigate hysteresis in stream water temperatures. One common cause of seasonal hysteresis is the influx of cold rain or snow melt in the spring, which results in spring water temperatures being lower than fall water temperatures at the same air temperature (Webb and Nobilis, 1997). As shown in Figure 2.5, the estimated varying intercept in the full VCM is larger in fall than that in spring. It indicates that the mean maximum water temperature in fall is higher than the mean maximum water temperature in spring. Therefore, the full VCM gives clear evidence for the presence of hysteresis at Site 5.

For the semi-VCM, the intercept θ_0 represents a global mean daily maximum water temperature and the varying slope $\theta_1(t)$ tends to measure the ratio of the water temperature to air temperature given the mean maximum water temperature for the whole year. In Figure 2.5, the intercept of the semi-VCM is very close to the intercept of the linear regression model, which is the mean maximum water temperature for the year for the centered data. Because semi-VCM uses a global intercept, its slope coefficient varies around the linear regression slope. Thus the semi-VCM provides information about the global sensitivity of water temperature to changes in air temperature and how it varies over time.

To further demonstrate the properties of the semi-VCM, Figure 2.6 plots the intercepts and

varying slopes of the semi-VCM for the 10 sites. It is clear that the values of 10 intercepts are all between 10^oC and 15^oC. The slopes for the ten sites tend to have a consistent pattern with a decline in spring and fall. The increased variability in air temperature results in a smaller slope in these periods. The range (the difference between the maximum and the minimum) of the slope in the semi-VCM relates to the variance of water temperature, and the location of maxima and the minima of the slope relates to the variance of the air temperature at different time points. So, by studying the slope curves in Figure 2.6, I obtain information about the similarity and variance in the water and air temperature relationship for all ten sites.

The full VCM captures the relationship on a level that is more local in time than the semi-VCM. Because of the shorter focus, the slopes tend to be smaller and smoother as there is less change in temperature over shorter time periods. As indicated in Figure 2.5, the majority of the variability is associated with the change in the intercept of the full VCM. This is also the reason for the improved fit of the full VCM for the data sets. Therefore, I suggest using the semi-VCM when the air and water variations are small and consider using the full VCM when the air and water variations are large.

2.5 Discussion

The development and evaluation of the VCM require some choices which are discussed further below. These include the selection of the number of knots, selection of smoothing parameter, and the evaluation of the VCM relative to the logistic model. The advantages of VCM and future work are also included in this section.

Chapter 2. A Varying Coefficient Model for Single Stream.: *Li H, Deng X, Kim D-Y, Smith E. P, 2014. Modeling maximum daily temperature using a varying coefficient regression model. Water Resources Research 50: 3073-3087.*

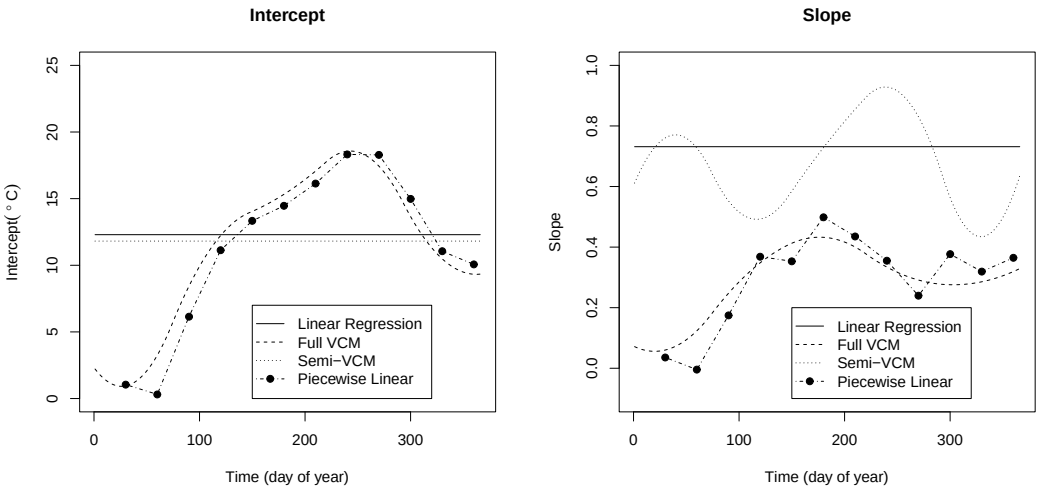

Figure 2.5: The intercept and the slope plot for different models for Site 5.

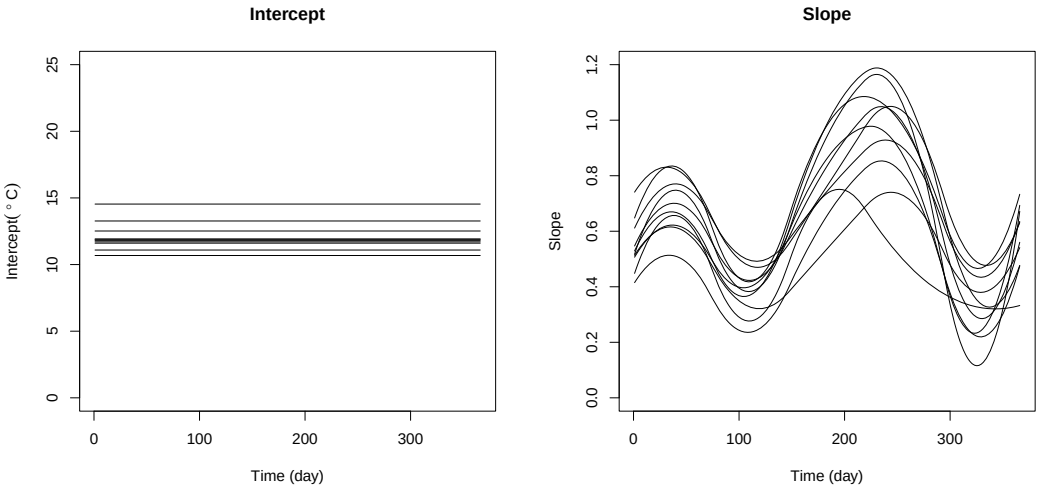

Figure 2.6: The intercept and the slope plot in semi-VCM for all the 10 sites.

Chapter 2. A Varying Coefficient Model for Single Stream.: *Li H, Deng X, Kim D-Y, Smith E. P, 2014. Modeling maximum daily temperature using a varying coefficient regression model. Water Resources Research 50: 3073-3087.*

2.5.1 Smoothing Parameter Selection

In Section 2.3.2, I introduced LOOCV and GCV as methods to select optimal smoothing parameter λ. There are other criteria that might be used, such as maximum likelihood (ML), restricted maximum likelihood (REML) and Mallow's C_p statistic (Ruppert *et al.*, 2003). ML and REML are likelihood-based approach and are not appropriate for the case here because the VCM in this work does not rely on an assumed distribution. GCV is approximately equal to C_p and does not require a prior estimate of the variance of the error term (Ruppert *et al.*, 2003). Therefore, I chose GCV as the criterion for smoothing parameter selection.

2.5.2 Knot Selection

Besides the tuning parameter λ in (2.6), the number of knots also affects the smoothness of the varying coefficient curves. What is more, the number of knots determines the number of parameters in the VCM. In section 2.4.1, I fixed the number of knots at $N = 4$ because it provides both smooth varying coefficient curves and high NSC statistics. Two commonly used criteria for model selection are the Akaike Information Criteria (AIC) and Mallow's C_p (Ruppert *et al.*, 2003). The AIC is a likelihood-based criterion and is not appropriate here. An approach based on a statistic such as C_p could be used as I focus more on smoothness of the coefficient curves and NSC as criteria for model selection. As part of a sensitivity analysis I compared fits using $N = 3$ and $N = 5$ knots as well as different degrees of polynomials for the spline models and chose to use N=4 knots with quadratic splines as these choices resulted in smooth curves, parsimony and good cross-validation statistics.

2.5.3 Hypothesis Testing

In this work, I developed a block bootstrap hypothesis testing procedure for examining the VCM relative to the linear regression model. However, I have not constructed a testing procedure to compare the VCM with the nonlinear logistic model. Note that the logistic model is not a nested model of VCM. The usual hypothesis testing approaches, such as those based on the F-test (Rencher and Schaalje, 2008), might not be applicable. To evaluate the effect of using the VCM when the underlying model is the logistic model, I conducted a simulation study. Specifically, using air temperatures from Site 5, water temperatures are simulated based on the logistic function ($\mu = 0, \alpha = 29.2, \beta = 12.9, \gamma = 0.21$) with standard normal errors. Because water temperature could not be below zero, negative values in the simulated data are truncated to be zero for water temperature. I generated 1000 simulation data sets and applied the two models to each data set. The average NSC from the nonlinear logistic model was 0.97 and the average NSC from full VCM was 0.95. The results show that the VCM is comparable to the logistic model even when the true model is the nonlinear logistic model.

2.5.4 Advantages of VCM over Linear Model

It is worthwhile to note that for some sites, the linear regression model could fit the data quite well (for example, NSC statistics for site 6 was 0.91). Even so, the proposed VCM method still has several merits compared with the linear regression model. First, the linear regression model is a special, degenerate case of the VCM. One could apply the goodness-of-fit test in section 2.3 to check whether the VCM is significantly different from the linear regression model. Moreover, by dynamically incorporating temporal information into the model, the VCM will always improve the prediction accuracy of the estimated model. In

addition, the VCM provides a meaningful interpretation of the variation in the water-air temperature relationship over time. For sites with high NSCs, the relationship between air and water temperatures might vary due to various effects such as seasonal hysteresis. The VCM can automatically account for such variation in the air and water relationship across time.

2.6 Conclusions

I developed time-varying coefficient models for studying the relationship between daily maximum water and air temperatures for 10 stream sites in Maryland, West Virginia, Virginia, North Carolina and Georgia. Statistical inferences using bootstrap hypothesis testing were also developed to examine the appropriateness of the proposed models. The proposed method effectively quantifies the water-air temperature relationship allowing for flexibility in local or global trend. Both the proposed full VCM and semi-VCM result in reasonably accurate prediction for the data from all 10 sites, having lower RMSEs than the linear regression model and the nonlinear logistic model. Moreover, the proposed models provide meaningful interpretations of the temporally dynamic relationship between air and water temperature.

The VCM is superior for these data sets as the effect of seasonal hysteresis is a significant determinant of water temperature. Mohseni *et al.* (1998) and Mohseni *et al.* (1999) proposed separating the annual cycle into periods according to the warming season and the cooling season, and suggested analyzing the air-water temperature relationship separately as a way to address the problem. The full VCM is able to capture seasonal dynamics in water and air temperatures without having to separate data into different time intervals. It thus automatically accounts for the hysteresis in the streams using time varying coefficients.

Chapter 3

Stream Clustering Based on Water-Air Relationship

3.1 Introduction

There are two major challenges for clustering streams based on water-air temperature relationship. The first challenge is to work with bivariate curves associated with the air and water temperatures. Because each site has two curves, the observations are not vectors but matrices. Hence creating an interpretable distance measure is not trivial. Ieva *et al.* (2013) adopt a classical distance approach by summing distances between curves for electrocardiography (ECG) signals. In a case study on air temperature and precipitation in Canada, Jacques and Preda (2014) represent multivariate curves through principal components, which are defined based on the projections of functional data on eigenfunctions. They put equal weight on each curve and sum the projected distances between curves. Such a method may not be appropriate when one of the curves is of more interest or has greater significance to the clustering procedure. The second challenge is to incorporate spatial information from the streams into

the clustering method. Streams used to collect water and air temperatures are spatially located within a fixed geographical region and hence are likely to be correlated. Therefore, it is meaningful to incorporate spatial correlation into the clustering method. Several researchers have worked on spatially adjusted distance and/or the clustering algorithm based on spatial correlation. Oliver and Webster (1989) modify curve distances through the variogram (Cressie, 1993). Giraldo *et al.* (2012) extend Oliver and Webster (1989)'s method and apply hierarchical clustering to spatially correlated data. Ben-Dor *et al.* (1999) propose a cluster affinity search technique (CAST) for spatially constrained clustering. Brenden *et al.* (2008) modify the CAST method and propose a valley segment affinity search technique to cluster streams and sites within streams. However, these cluster algorithms might not be easily extended to bivariate functional data without defining an appropriate distance measure.

In this chapter, I propose a weighted distance measure for clustering based on the water-air temperature relationship to deal with the first challenge. Specifically, I cluster streams based on profiles with two curves: the time varying average daily maximum water temperature and the sensitivity of water temperature to the changes in air temperature. The varying coefficient linear model (VCM) (Li *et al.*, 2014) is used to quantify the water-air temperature relationship for individual streams. The intercept and slope functional curves obtained from the VCM correspond to the smoothed average water temperature and the sensitivity of water temperature to air temperature over time. I then develop a weighted Canberra distance (Lance and Williams, 1967) to measure the distance between streams based on the intercept and slope curves. The weighted distance provides flexibility to enable the resultant clustering to balance the focus between the two curves. That is, different weights lead to different meaningful interpretations associated with the clustering results. More weight on the intercept curves will result in clusters of streams separated mainly by smoothed water temperature. In contrast, more weight on the slope curves will result in clusters of streams

separated mainly by sensitivity of water temperature to air temperature.

To deal with the second challenge of incorporating spatial information into clustering, I adopt the idea in Giraldo *et al.* (2012) to adjust the weighted distance by a scale of spatial distance using the variogram (Cressie, 1993). The proposed distance measure is the product of the weighted distance between the two curves in the VCM and the variogram. This model-based spatially adjusted distance makes clusters both statistically interpretable and spatially meaningful. The model-based weighted distance focuses on the water-air temperature relationship and the spatial component leads to streams having similar climate and landscape features within clusters. The K-medoids clustering algorithm is applied to this spatially adjusted weighted distance.

It is worth noting that the interpretability of the cluster analysis results often depends on the choice of distance measure and the clustering algorithm. For the proposed method, the flexibility in the weighted and spatially correlated distance measure leads to meaningful interpretations of grouped streams. The weight parameter can be used to vary the emphasis from an analysis on water temperature to an analysis of the water-air relationship (sensitivity). Moreover, such flexibility will help reveal several common characteristics for different stream groups, such as the importance of solar radiation and percent forest coverage, under various weighted distance measures. The parameters in the variogram can control how much spatial correlation is incorporated into the analysis. The proposed method not only easily balances statistical variation (water-air temperature) and spatial variation (location of streams), but also will be helpful for natural resource management strategies.

3.2 Data

In this chapter, 62 streams with complete records were used; their locations are shown in Figure 3.1. For each site, I extract ten months of data with the same starting date (1 January 2011) and ending date (15 October 2011). In this study, the *daily maximum values* are used as these values representing upper temperatures that the fish are exposed to during the period. The data used as inputs consist of $T = 288$ paired daily maximum air and water temperatures for each site.

For stream site $i = 1, 2, \cdots, n$ ($n = 62$), I denote daily maximum air and water temperature as $\mathbf{X}_i(t) = (\mathbf{A}_i(t), \mathbf{W}_i(t))$, where $\mathbf{A}_i(t)$ is the air temperature at time t, $\mathbf{W}_i(t)$ is the water temperature at time t and $t = 1, 2, \cdots, T$. Without loss of generality, I used centered air temperature. For each site, in addition to air and water temperature there are several stream site characteristics that were measured. One of them is a climatic variable, solar radiation, and four are landscape variables including latitude, longitude, elevation and the percentage of forest in the area around the stream. Those variables may affect the water-air relationship and have a connection with landscape management strategies (Mayer, 2012). For site i, denote $s_i = (u_i, v_i)$ as the location variable, where u is latitude and v is longitude. In our clustering methods, only latitude and longitude are used in the clustering algorithm along with water and air temperatures. Other covariates serve as descriptive information to evaluate clustering results.

3.3 Proposed Clustering Method

In this section, I detail the proposed bivariate functional data clustering method. It has the following three major steps. I first apply the VCM to the original water and air temperature

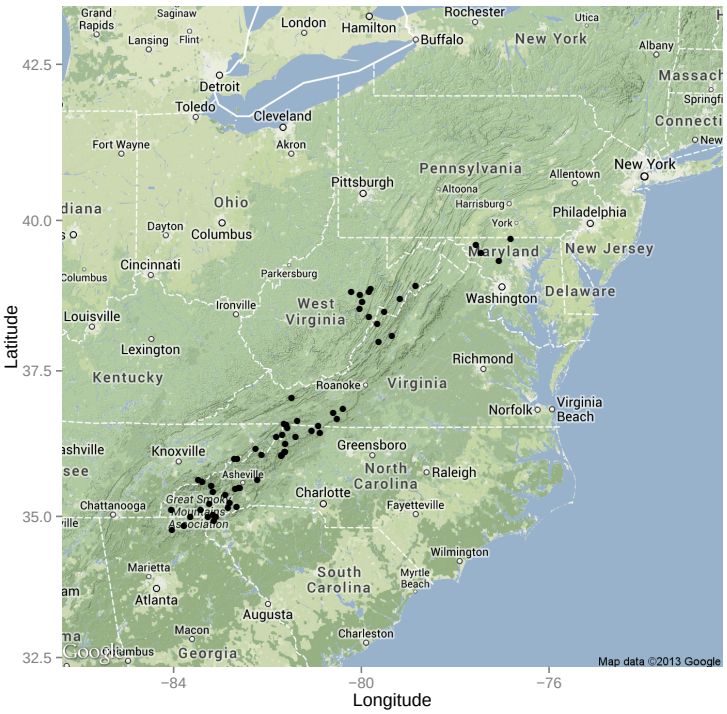

Figure 3.1: Location of 62 streams (black dots are the locations).

data, and obtain the bivariate functional curves to describe the water-air relationship. Based on the bivariate functional data, I then calculate a spatially-adjusted weighted distance measure, which incorporates the latitude and longitude through the variogram. Using the weighted distance measure, I finally apply the K-medoids method (Hastie *et al.*, 2009) to implement the clustering algorithm.

3.3.1 VCM for Water-Air Relationship

It is common to smooth temporal data before clustering such that the curves observed at discrete time points will have a continuous functional form. (Ramsay and Silverman, 2005). Popular smoothing techniques include the Karhunen-Loève expansion (Chiou and Li, 2007), the orthonormalized Gaussian basis (Kayano *et al.*, 2010) and the B-spline (Abraham *et al.*, 2003) approaches. In this work, due to the focus of the study, I need a smoothing method to model the sensitivity of water temperature to the change of air temperature. Since the VCM provides a meaningful interpretation of the water and air relationship, it has advantages over other approaches to fit the study goal. In this work, I use the VCM as the smoothing method. Specifically, the intercept profile obtained from the VCM provides information about the smoothed maximum water temperature and the slope profile reflects the sensitivity of water temperature to changes of air temperature over time. By using varying coefficient profiles as functional data, the clustering results can be expected to produce groups with similar water and air temperature relationships.

For site i, I consider the varying coefficient model (Li *et al.*, 2014) for the air-water temperature relationship as

$$W_i(t) = \theta_{0i}(t) + A_i(t)\theta_{1i}(t) + \epsilon_i(t), \tag{3.1}$$

where $\theta_{0i}(t) = \sum_{j=1}^{K} \alpha_{ij} b_{ij}(t)$ and $\theta_{1i}(t) = \sum_{j=1}^{K} \beta_{ij} b_{ij}(t)$ are varying intercept and slope

coefficients. The $\epsilon_i(t)$ is the error term in the model with $E(\epsilon_i(t)) = 0$ and $\text{var}(\epsilon_i(t)) = \sigma_i^2(t)$. Here $\{b_{i1}(t), \cdots, b_{iK}(t)\}$ is a set of K basis functions and $\alpha_{ij}, \beta_{ij},\ j = 1, 2, \cdots, K$, are parameters in the VCM in (3.1). To estimate these parameters, I adopt the regression spline methods used in Hoover *et al.* (1998) and Li *et al.* (2014). For each site i, the estimated coefficients $\mathbf{X}_i(t) = (\hat{\theta}_{0i}(t), \hat{\theta}_{1i}(t))$ are the profiles of the bivariate curves used in the following analysis.

3.3.2 Distance Measure

In this section, I develop a weighted distance measure for bivariate functional data and incorporate spatial correlation between stream sites into this distance. The general form of the proposed distance between site i and site j can be written as

$$d_s(i, j) = r(h)d(\mathbf{X}_i(t), \mathbf{X}_j(t)), \tag{3.2}$$

where $h = \|s_i - s_j\|$ is the Euclidean distance between sites i and j. Here $d(\mathbf{X}_i(t), \mathbf{X}_j(t))$ is used to quantify the distance between two sites based on the bivariate functional profile (i.e., the relationship between water-air temperatures) and $r(h)$ is a function of the spatial distance between two sites based on longitude and latitude.

Note that the functional profile for each site includes an intercept curve and a slope curve. Therefore, the calculation of $d(\mathbf{X}_i(t), \mathbf{X}_j(t))$ would involve two distances, the distance between the intercept curves and the distance between the slope curves. To combine the two distances, I consider a weighting scheme to create a single measure, where the weight can be chosen to give different emphasis on slope and intercept. Specifically, I define

$$d(\mathbf{X}_i(t), \mathbf{X}_j(t)) = wd_c(\theta_{i0}(t), \theta_{j0}(t)) + (1 - w)d_c(\theta_{i1}(t), \theta_{j1}(t)), \tag{3.3}$$

where $0 \leq w \leq 1$ is the weight, and $d_c(\cdot, \cdot)$ is a Canberra distance (Lance and Williams, 1967) defined as

$$d_c(\theta_{ik}(t), \theta_{jk}(t)) = \int \frac{|\theta_{ik}(t) - \theta_{jk}(t)|}{|\theta_{ik}(t)| + |\theta_{jk}(t)|} dt, \tag{3.4}$$

for two curves $\theta_{ik}(t)$ and $\theta_{jk}(t)$, $k = 0, 1$. Because the Canberra distance is standardized, the two distances may be combined into a a single metric and I avoid the problem of different magnitudes associated with the intercept and the slope curves. The weight, w, here provides flexibility for clustering streams based on different characteristics. For example, when w is close to 0, the distance in equation (3.3) contains more information on the sensitivity of water temperature to air temperature (slope curve). Then the resultant clusters tend to separate streams based on water-to-air temperature sensitivity. In contrast, when w is close to 1, the distance in equation (3.3) contains more information on smoothed maximum water temperature (intercept curve). Then the smoothed maximum water temperature will become a key factor in clustering. Therefore, I can alter the emphasis of the cluster results based on different weights.

For the spatial distance $r(h)$, I consider using the variogram, which is discussed in Oliver and Webster (1989), Cressie (1993) and Giraldo *et al.* (2012). Specifically, I adopt the distance measure in Giraldo *et al.* (2012) and extend it to bivariate curves. As there are two curves for each site, $\mathbf{X}_i(t) - \mathbf{X}_j(t)$ is not a vector. Therefore $|\mathbf{X}_i(t) - \mathbf{X}_j(t)|$ is more appropriate defined by using the distance measure in equation (3.3) (i.e., $|\mathbf{X}_i(t) - \mathbf{X}_j(t)| = d(\mathbf{X}_i(t), \mathbf{X}_j(t))$) in the variogram for bivariate functional data. For the form of variogram, I use the method suggested by Oliver and Webster (1989) and express the variogram as

$$r(h) = c_0 + c_1(1 - exp\{-\frac{|h|}{\rho}\}). \tag{3.5}$$

To extend the exponential function in Oliver and Webster (1989), I added parameters c_0

and c_1, where c_0 is nugget effect and $c_0 + c_1$ is the sill (Le and Zidek, 2006). By plugging equation (3.5) into (3.2) and dividing by $c_0 + c_1$, I obtain the following expression, which is also used in Oliver and Webster (1989):

$$d^*(i,j) = \frac{c_0}{c_0+c_1} d(\mathbf{X}_i(t), \mathbf{X}_j(t)) + \frac{c_1}{c_0+c_1}(1 - exp\{-\frac{\|\mathbf{h}\|}{\rho}\}) d(\mathbf{X}_i(t), \mathbf{X}_j(t)). \tag{3.6}$$

The first part of equation (3.6) is proportional to the distance (3.3) and the second part can be modified according to the spatial correlation. The distance measure (3.6) is the distance I will use in the clustering algorithm.

To estimate the parameters in the variogram in distance measure (3.6), I adopt a classical formula introduced in Cressie (1993):

$$2\hat{\gamma}(h) \equiv \frac{1}{N_h(N_h-1)} \sum_i \sum_j ((\mathbf{X}_i(t) - \mathbf{X}_j(t))^2 : (i,j) \in N(h); h \in T(h(l))), \tag{3.7}$$

where

$$N(h) \equiv \{(i,j) : s_i - s_j \in (h-\epsilon, h+\epsilon); i,j = 1,2,\cdots,n\},$$

N_h is the number of sites in $N(h)$, $T(h(l))$ is the l^{th} tolerance region, $l = 1,2,\cdots,m$ and m is the number of tolerance regions. In selecting tolerance regions, I keep sizes of regions similar and make the number of distinct pairs to be at least 30 (Journel and Huijbregts, 1978). For bivariate data, I defined the empirical estimate of the variogram as

$$2\hat{\gamma}(h(l)) \equiv \frac{1}{N_h(N_h-1)} \sum_i \sum_j (d(\mathbf{X}_i(t), \mathbf{X}_j(t))^2 : (i,j) \in N(h); h \in T(h(l))), \tag{3.8}$$

To estimate c_0, c_1 and ρ in (3.5), I use nonlinear least squares regression between the empirical variogram in (3.8) and the parametric variogram in (3.5).

3.3.3 Clustering Algorithm

There are various methods used in functional clustering such as K-means and hierarchical clustering (Hastie *et al.*, 2009). Sangalli *et al.* (2010) consider the misalignment of functional data and define similarity between curves in order to apply K-means. Tokushige *et al.* (2007) extend the K-means method for functional data and propose crisp and fuzzy K-means algorithms. Tarpey and Kinateder (2003) propose a K-means method based on principal points. Giraldo *et al.* (2012) apply hierarchical clustering method to spatially correlated data in Canada. Classical K-means requires squared Euclidean distance, which is not applicable to our proposed distance. Alternatively, K-medoids clustering (Hastie *et al.*, 2009) can be more flexible for different distance measures.

Regarding the choice of the number of clusters k, it often depends on prior information or the goal of study. In this work, I expect the number of clusters to be relatively small for meaningful interpretation. A naive method to determine the number of clusters k is through visualization of the stream temperature series. Such an approach is subjective but helpful when there is a clear pattern of separate groups. In our study, in which I use a spatially weighted distance measure, visualizing curves on a graph can be difficult. Alternatively, I use an analytical method to determine k based on the calculation of within cluster dissimilarity. Such an idea is also used in Tibshirani *et al.* (2001). They compare logarithms of within cluster dissimilarity of original data to uniformly generated data and estimate the optimal k by the gap between the two. I adopt a similar but simpler method by selecting k based on the silhouette width (Rousseeuw, 1987). Specifically, for a clustering with k clusters $C_1, \ldots, C_k$, the silhouette width for stream i is defined as

$$s(i|C_1, \ldots, C_k) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \tag{3.9}$$

where $a(i) = \frac{1}{|C(i)|} \sum_{j \neq i, j \in C(i)} d_s(i,j)$ is the average distance of stream i to all other streams within the same cluster. Here $C(i)$ is defined as the cluster that stream i belongs to. Also $b(i) = \min_{i \notin C_h} \frac{1}{|C_h|} \sum_{j \in C(h)} d(i,j)$ is the minimum average distance from stream i to all points in another cluster. Then the optimal number of clusters k is determined by

$$k = \arg\max_m \sum_i s(i|C_1, \ldots, C_m) \tag{3.10}$$

3.4 Results

In this section, 62 streams with complete ten-month water and air temperature data are used in the analysis. I fit the VCM to the paired temperature data for each stream and obtain the intercept and slope curves. The proposed clustering method is applied to bivariate curves of intercept and slope. For illustration, the performance of the proposed method is evaluated by choosing various values of weight w in equation (3.3), i.e., $w = 100\%, 75\%, 50\%, 25\%$ and 0%. In the following sections, the clustering results will be viewed from both statistical and geographical perspectives.

3.4.1 Parameter Estimation

The VCM method in Section 3.3 is used to fit the water and air temperature data of each stream in order to quantify the water-air temperature relationship. The fitted coefficient curves are shown in Figure 3.2. For the intercept curves in Figure 3.2, there is no visible clustering in the pattern of smoothed maximum water temperature trend for the streams. The magnitude of the smoothed maximum water temperature varies considerably: for each time point, the gap between the maximum intercept and the minimum intercept is about $7°C$. For the slope curves in Figure 3.2, the variation is relatively large, especially in the

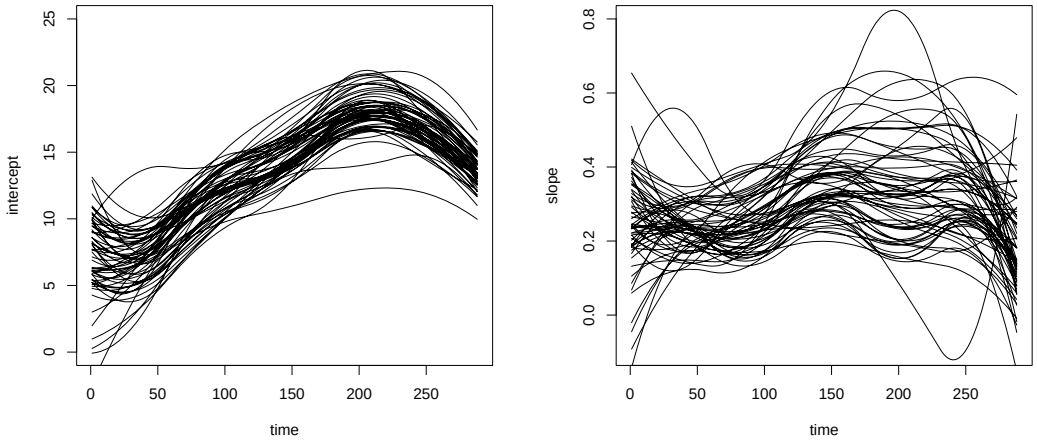

Figure 3.2: Intercept and slope curves from VCM for 62 sites.

summer time (around the period of $t = 200$). Recall that the intercept and slope curves represent the smoothed maximum water temperatures and sensitivities of daily maximum water temperature to daily maximum air temperature, respectively. It is thus helpful to use the bivariate intercept and slope curves obtained from the VCM for clustering streams and identifying streams with high risk to increasing air temperature.

In order to estimate the variogram, the pre-specified formulation in equation (3.5) is used. Figure 3.3 shows the estimated variogram for five different values of w in (3.3) for the distance measure. From Figure 3.3, I consistently observe that the value of the variogram increases as the distance between two streams increases. Such an observation further verifies the fact that streams located closely have similar water and air temperature relationship. Therefore, it is reasonable to incorporate the spatial information into the clustering method.

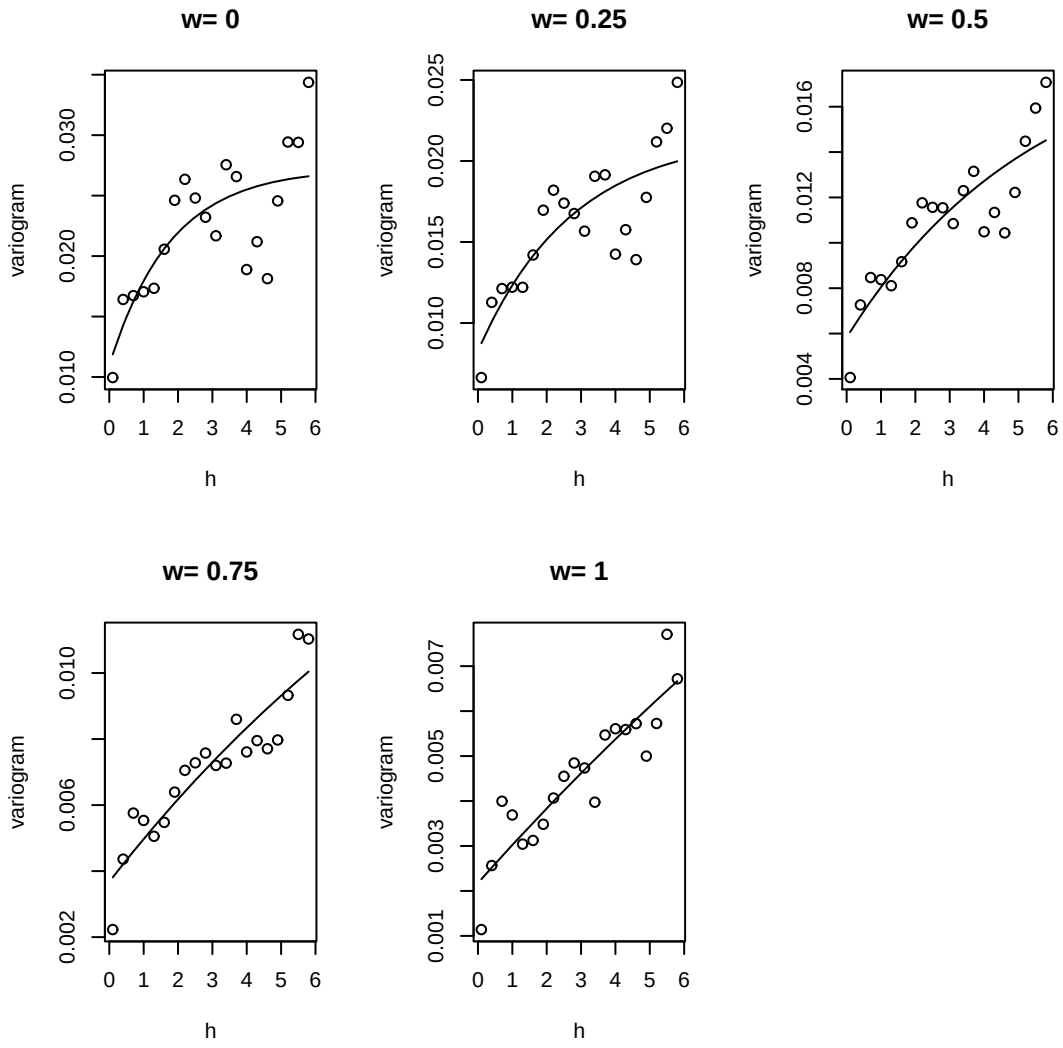

Figure 3.3: Variogram estimates and parametric form for different weights based on the intercept curve: circles are empirical estimates of the variogram and curves are for the estimated exponential variogram.

3.4.2 Clustering Results

The silhouette statistics for choosing the number of clusters are summarized in Figure 3.4. From the results in Figure 3.4, it clearly shows that two or three clusters can be the optimal number of clusters, producing the highest silhouette statistics for all scenarios of five weights. Figures 3.5-3.9 report the performance of the proposed clustering method with respect to the weights, $w = 100\%, 75\%, 50\%, 25\%$ and 0%, on the intercept. For each figure, part (a)-(f) display the location of streams, intercept and slope curves, boxplots for solar radiation, percent forest, and elevation, respectively. For the streams in each cluster, I use their solar radiation, percent forest and elevation to form separate boxplots, respectively. By changing weight w in distance measure (3.3), I have three interesting findings.

First, the weight can determine the impact of spatial correlation on the cluster results. From Figure 3.5(a) where the weight on the intercept is large, I observe that the proposed clustering method clearly results in two groups of northern and southern streams on the map. For the other figures (part (a) of Figure 3.6-3.9) of which the weight for intercept decreases, the northern and southern clusters pattern is not that visible. Such a finding indicates that the major variation in intercept curves is related to geographical variation as the clustering results mainly depend on the location of streams. Since the intercept curve from the VCM is interpreted as the smoothed maximum water temperature, it is expected that streams with smaller mean intercepts are located in the north because streams with higher latitude often have lower stream temperature. In contrast, Figure 3.9(a) reveals that when the weight on the slope curve is large (w is small), the geographical pattern of stream groups becomes weaker. As shown in Figure 3.9(a), where $w = 0$ (distance is calculated based on slope only), the stream clusters are not aligned by location on the map. This finding indicates that the spatial correlation amongst the slope curves is weak.

Second, the intercept and slope curves have connections with the landscape and climate variables in the clustering results. From Figure 3.5, I observe that when more weight is given to the intercept in the distance measure (w is large), the resultant stream clusters are grouped in a consistent manner with stream elevations. From Figure 3.5(f), it is seen that the two clusters result in strong separation with respect to elevation. This result is expected since streams with high elevation usually have lower summer water temperature (Figure 3.5 (b) red curves) which is directly related to the intercept curves. It can also be seen from Figure 3.10 that the averaged intercept curve for cluster 2 (red curve, high elevation and low latitude) has lower summer temperature and higher winter temperature. It is evidence that the summer average maximum water temperature is associated with elevation and winter average maximum temperature is associated with latitude. Note that the elevation information is not used in the proposed clustering algorithm. This finding further confirms that the proposed distance measure defined in equation (3.2) is meaningful and adequate for studying the connection between the landscape variables and the clustering results. Moreover, I can see from Figure 3.9 that when the slope is emphasized in calculating the distance measure (w is small), streams in different clusters tend to have distinct magnitude for variables such as solar radiation and percent forest. For example, the boxplots in Figures 3.9(d) and 3.9(e) show that the means of solar radiation and percent forest are separated for different clusters. Since the slope is interpreted as measuring sensitivity, our findings could imply that the sensitivity of the water temperature to the air temperature is related to solar radiation and percent forest. From Figure 3.9(d), it is observed that for sites having less solar radiation (lower boxplot), water is less sensitive to air temperature (red curves in Figure 3.9(c)). Furthermore, from Figure 3.9(e), I can see that for the sites with higher percent forest (upper boxplot), water temperature is less sensitive to air temperature (red curves in Figure 3.9(c)) as expected. Therefore, the clustering analysis infers that solar radiation and percent forest actually affect the sensitivity of the water temperature to the air temperature.

Note that the slope curves for the second cluster (red curves) are flat during summer time (day 174 through 230) in Figure 3.9 part (c). Days 174 and 230 are actually two knots in our regression splines. Therefore I can take advantage of the VCM and test if the slope curves in cluster 2 are constant. Specifically, I test if combinations of $\beta's$ in model (3.1) are zero. I obtained those combinations from the 31 sites in cluster 2 and conduct Hotelling's T^2 test (Rencher and Christensen, 2012) for constant slope. The test result shows that the slope curves in cluster 2 are constant for the summer period. In cluster 2, the percent forest is high. The constant slopes reveal that the forest can help reduce the effect on the water temperature from high air temperature. It is meaningful because when the air temperature is high in summer, the water temperature in cluster 2 would not increase fast as there is buffering from the forest.

Third, the proposed clustering methods with the distance measure (3.3) has more advantages if using considerable amount of weight, i.e., $w = 50\%$, on both intercept and slope. Figure 3.7 reports the clustering results for the weight $w = 50\%$ (equal weight on the intercept and slope). From the boxplots in part (d), (e) and (f) of Figure 3.7, elevation, solar radiation and percent forest exhibit significant differences across clusters. Since equal weight ($w = 50\%$) on the intercept and slope curves is used to calculate the distance measure (3.3), the clustering result therefore can reflect information from all three variables, elevation, solar radiation and percent forest in the clusters. Recall that those three variables are not used in calculating the proposed distance measure. Such an observation confirms that the proposed distance measure is flexible for investigating which underlying variables affect the water-air relationship and I can construct cluster results based on different climate and landscape variables.

In Figures 3.5-3.9 where the weight on intercept deceases from 100% to 0%, I find that elevation differences between clusters is less significant while solar radiation and percent

forest differences become more significant amongst clusters. To verify this, I calculate the F statistic (Rencher and Schaalje, 2008) to evaluate, for each of those three variables, whether the means are the same amongst clusters. The results are reported in Table 3.1. The results of the F-test show that solar radiation and percent forest have p-values that are less than 0.01 except when the value of weight $w = 100\%$. This is evidence that the separation on the two variables amongst clusters is consistent with the weight on the slope curve (sensitivity). Note in the last column in Table 3.1, the p-value for elevation is larger when there is more weight on the intercept (w is large). This is evidence that the elevation differences amongst clusters is consistent with the weight on intercept curve (maximum water temperature). This plausible outcome would hardly be discovered without changing the weight in distance measure (3.3). Depending on the research interest, one could increase the weight on the intercept to result in more elevation variation in clusters or increase the weight on slope to result in more solar radiation and percent forest separation between in clusters.

Table 3.1: F statistics and p-values (in parenthesis) from F-test for cluster differences

Weight on intercept	Solar radiation	Percent forest	Elevation
0	17.6(< 0.01)	18.5(< 0.01)	1.84(0.17)
0.25	17.2(< 0.01)	17.5(< 0.01)	2.82(0.09)
0.5	8.47(< 0.01)	9.98(< 0.01)	13.9(< 0.01)
0.75	9.29(< 0.01)	10.0(< 0.01)	13.9(< 0.01)
1	4.23(0.043)	2.98(0.08)	12.3(< 0.01)

3.5 Discussion

Other choices of the distance measure and clustering methods are discussed further below. The advantages of our algorithm and future work are also included in this section.

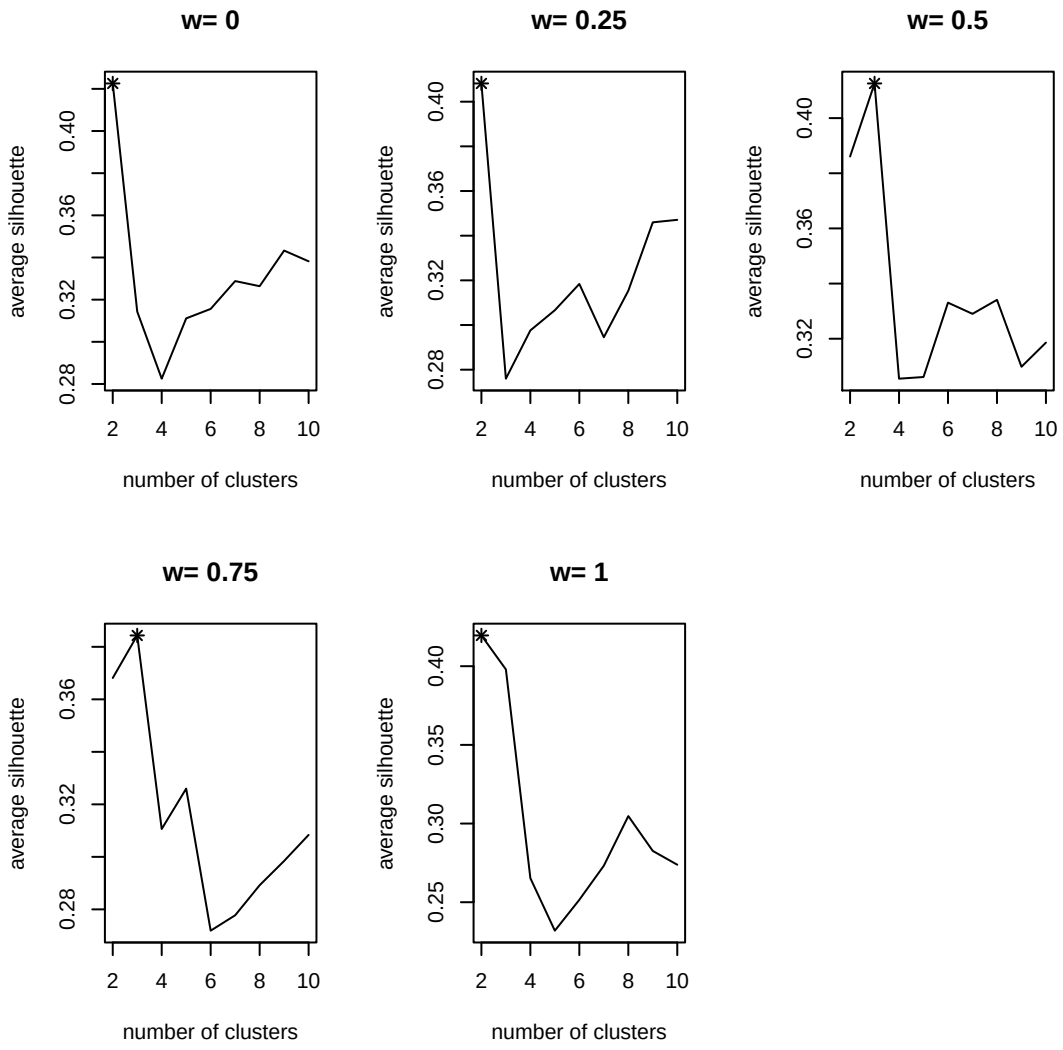

Figure 3.4: Silhouette statistics for different numbers of clusters under five different weights. The star marks indicate the optimal number of clusters and the corresponding Silhouette statistics.

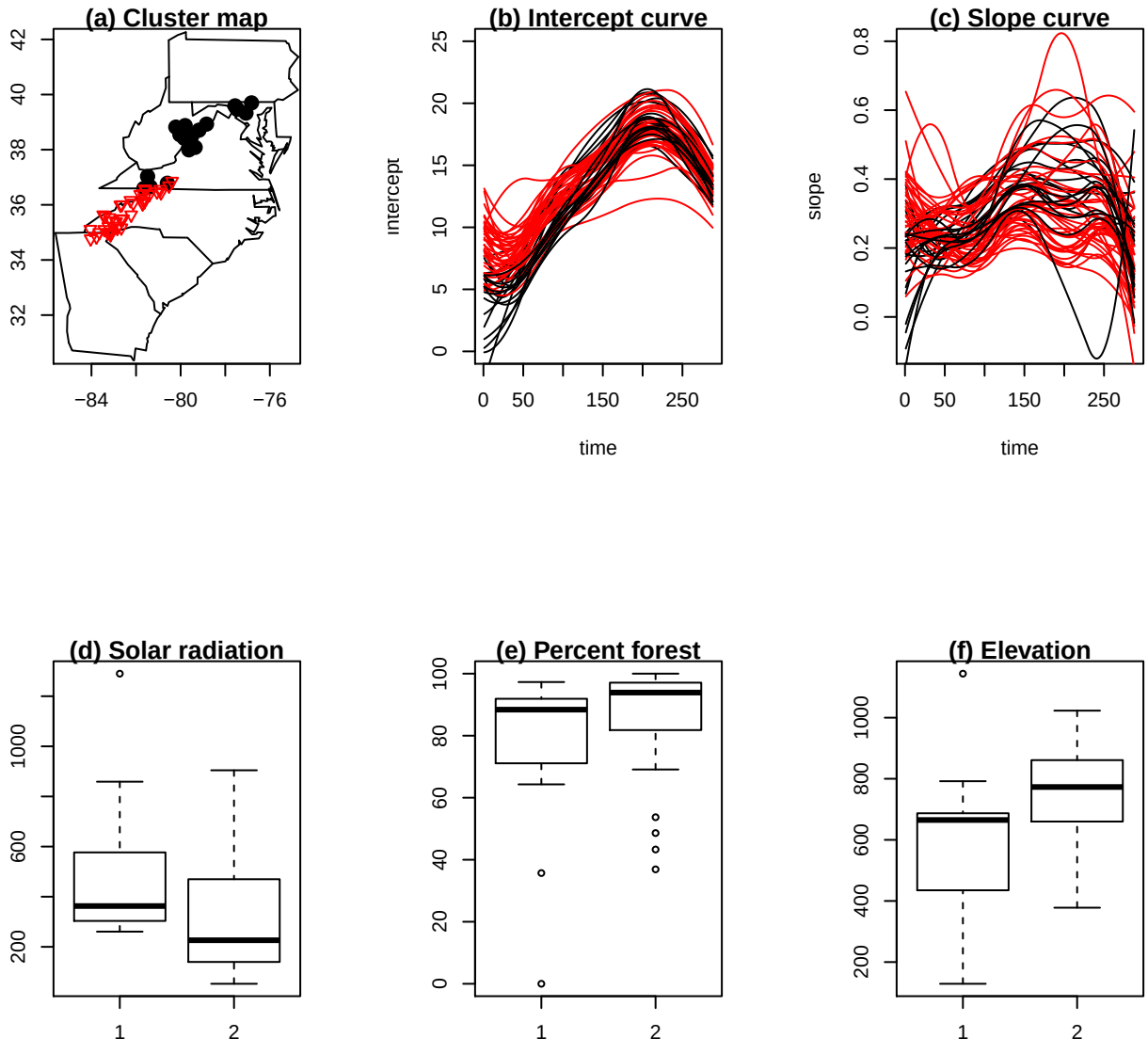

Figure 3.5: Cluster results for weight=100% on intercept. Numbers of sites in each cluster are: 21 for cluster 1 (black) and 41 for cluster 2 (red). (a) location of streams. (b) intercept curves. (c) slope curves. (d) boxplots for solar radiation for different clusters. (e) boxplots for percent forest for different clusters. (f) boxplots for elevation for different clusters.

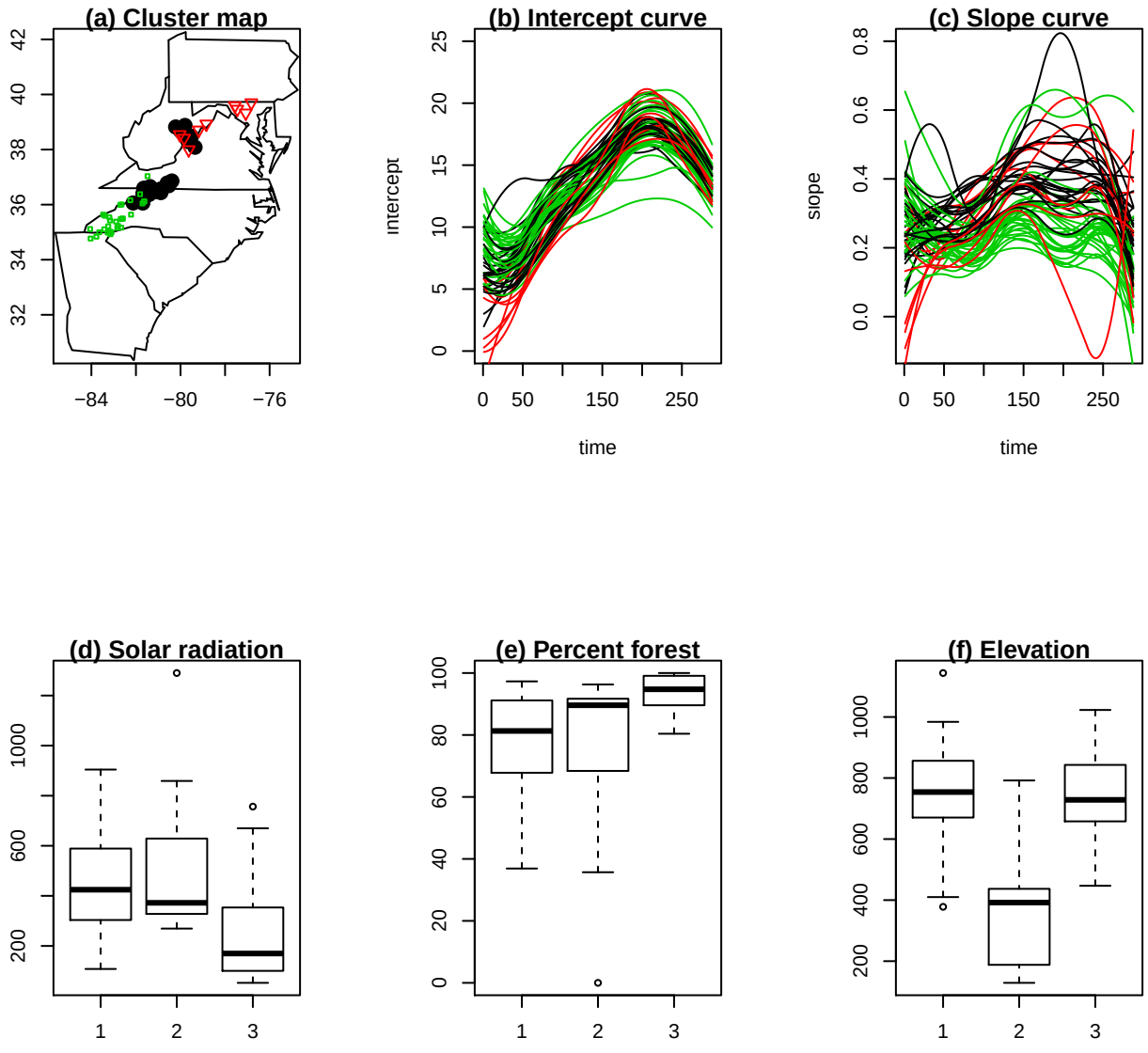

Figure 3.6: Cluster results for weight=75% on intercept. Numbers of sites in each cluster are: 23 for cluster 1 (black), 9 for cluster 2 (red) and 30 for cluster 3 (green). (a) location of streams. (b) intercept curves. (c) slope curves. (d) boxplots for solar radiation for different clusters. (e) boxplots for percent forest for different clusters. (f) boxplots for elevation for different clusters.

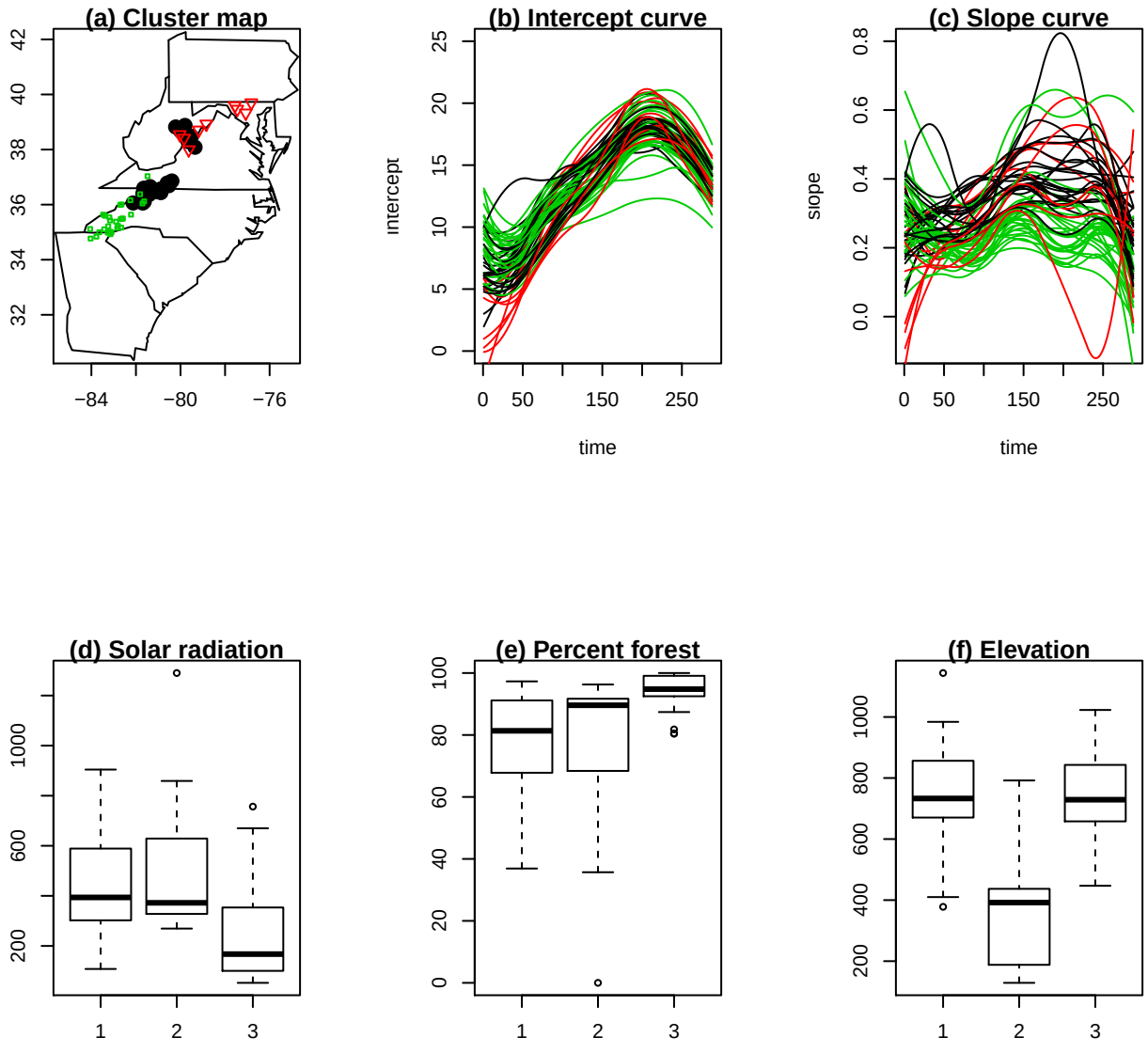

Figure 3.7: Cluster results for weight=50% on intercept. Numbers of sites in each cluster are: 24 for cluster 1 (black), 9 for cluster 2 (red) and 29 for cluster 3 (green). (a) location of streams. (b) intercept curves. (c) slope curves. (d) boxplots for solar radiation for different clusters. (e) boxplots for percent forest for different clusters. (f) boxplots for elevation for different clusters.

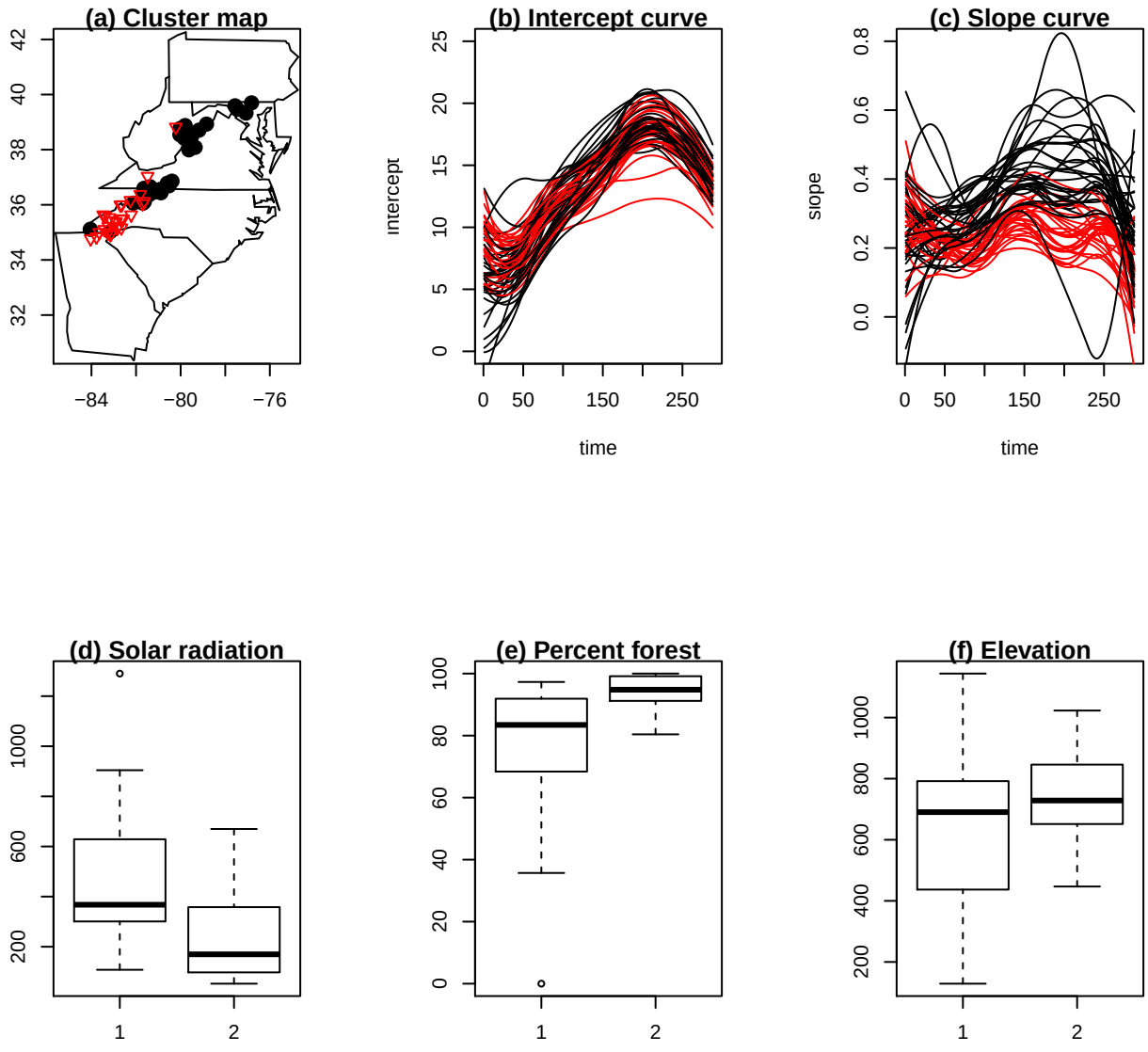

Figure 3.8: Cluster results for weight=25% on intercept. Numbers of sites in each cluster are: 34 for cluster 1 (black) and 28 for cluster 2 (red). (a) location of streams. (b) intercept curves. (c) slope curves. (d) boxplots for solar radiation for different clusters. (e) boxplots for percent forest for different clusters. (f) boxplots for elevation for different clusters.

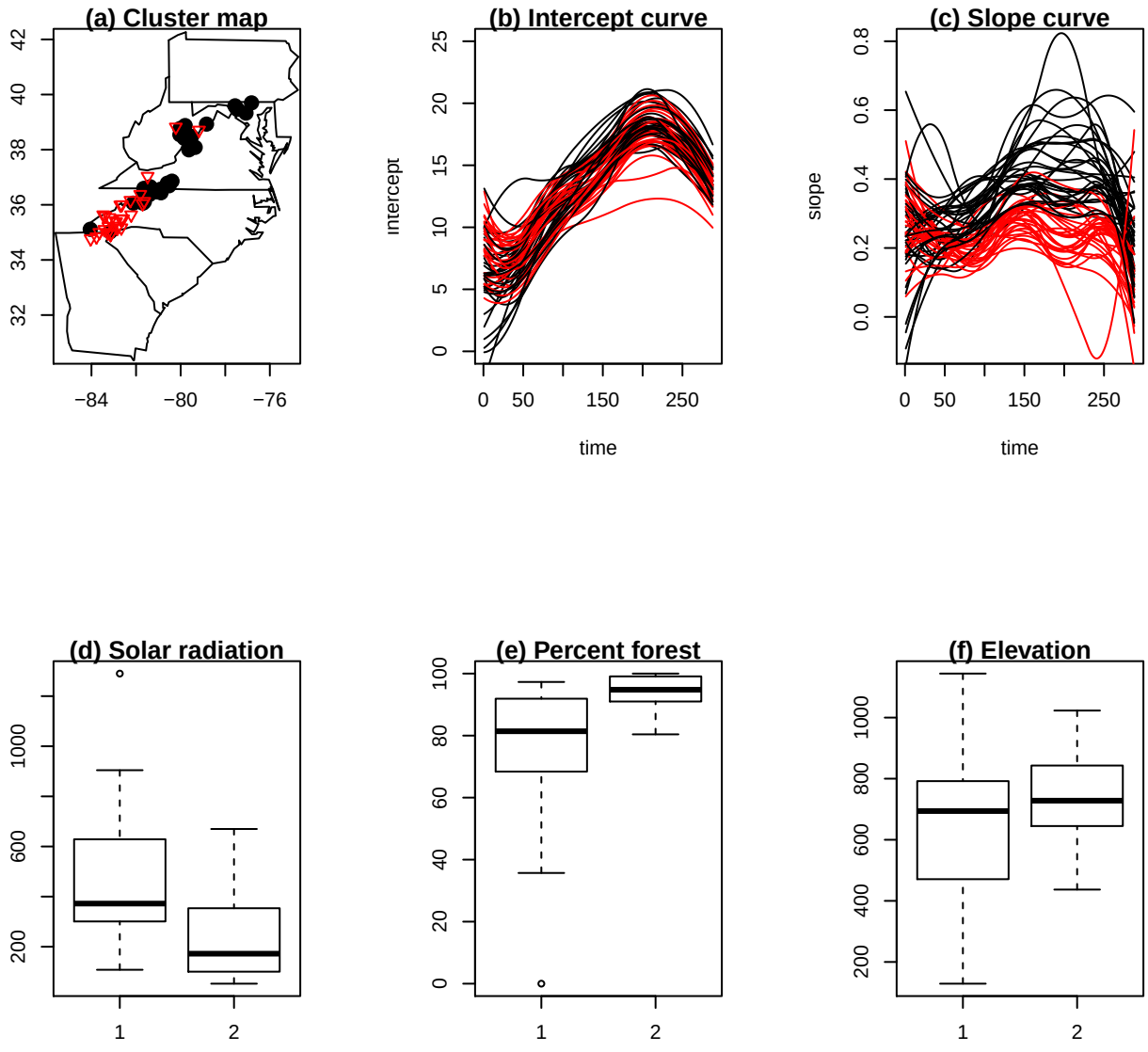

Figure 3.9: Cluster results for weight=0% on intercept. Numbers of sites in each cluster are: 33 for cluster 1 (black) and 29 for cluster 2 (red). (a) location of streams. (b) intercept curves. (c) slope curves. (d) boxplots for solar radiation for different clusters. (e) boxplots for percent forest for different clusters. (f) boxplots for elevation for different clusters.

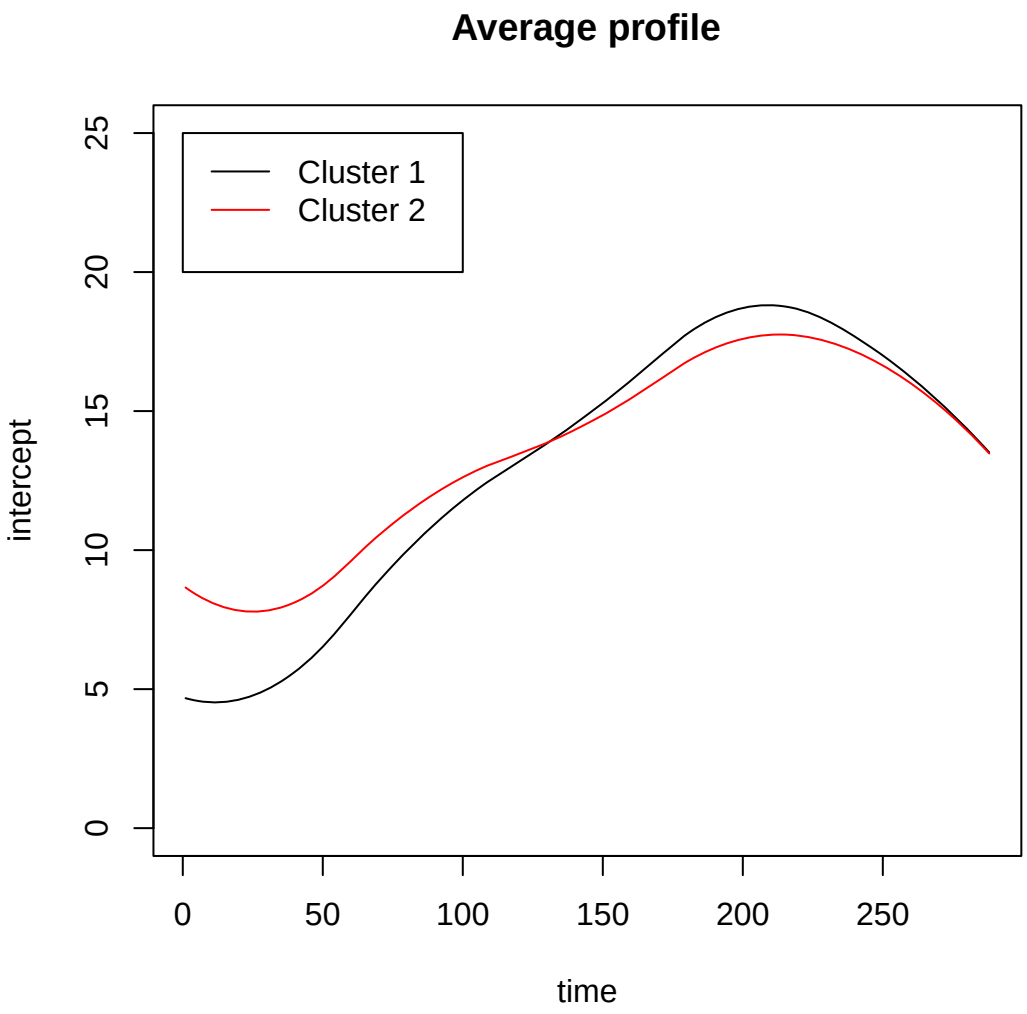

Figure 3.10: The averaged intercept curves for two clusters when w=1.

Another natural measure that incorporates spatial information can be the spatially penalized dissimilarity measure

$$d^*(i,j) = d(\mathbf{X}_i(t), \mathbf{X}_j(t)) + \lambda \|s_i - s_j\|, \tag{3.11}$$

where $\|.\|$ is the Euclidean distance and λ is a tuning parameter that balances distance between profiles and spatial distance. A good feature of (3.11) is that the value λ can be used to interpret clustering results. When $\lambda = 0$, the cluster result is based on curve distance and when λ is large, the cluster result emphasizes on spatial distance. I can see that distance (3.6) has a similar form with distance (3.11). If the nugget variance (c_0) is large, $d^*(i,j)$ gives more weight to variation between profiles, if c_1 is large, $d^*(i,j)$ puts more weights on the spatial variation.

In this chapter, I applied K-medoids clustering algorithm for the proposed distance measure in (3.3). Classical methods such as hierarchical clustering (Giraldo *et al.*, 2012) are often used in environmental functional data clustering. Both K-medoids and hierarchical clustering methods result in similar groups of streams for our data. I chose K-mediods method because it has two advantages. First, it is an iterative optimization procedure such that it can gradually improve the clustering quality. Second, the K-medoids algorithm is effective for detecting compact spherical-shaped clusters and is easy to use in practice (Aggarwal and Reddy, 2013).

3.6 Conclusion

In this chapter, I develop a bivariate functional clustering procedure to cluster multiple streams based on the water and air temperature relationship. A varying coefficient model is used to describe this relationship and provides meaningful interpretation of how the relationship changes over time. I define a weighted distance measure adjusted by spatial

information and apply the K-medoids clustering method using this distance. The data analysis shows that streams in same cluster share similar values of solar radiation, percent forest and elevation.

The challenge of clustering bivariate functional data is that there are two groups of observations associated with one unit (i.e., streams). The additivity assumption in Ieva *et al.* (2013) puts equal weight on multiple curves thus treats them equally. Although this approach reduces the number of curves, it lacks flexibility in interpreting the effect of each curve on clustering results. The weighted distance for two curves provides this flexibility and can be used for different research purposes. In this water and air temperature study, more weight on the intercept of VCM is seen to result in streams that are grouped by smoothed maximum water temperature and more weight on the slope will result in streams grouped by the water and air slope relationship.

Weighted distance can also be used in the variogram definition for bivariate functional data. Incorporating spatial correlation by variogram enables sites in the same cluster to have both similar water-air relationships and geographic characteristics. This is important in that the cluster results could provide information on how land management can be tailored to the water-air relationship. For example, in Figure 3.8 and 3.9, I found that sensitivity of water temperature to air temperature is associated with percent forest. The results suggest that the set of streams with high percent forest have buffered sensitivity to changes in air temperature. Specifically, the sites have a relatively constant relationship between air and water temperature, especially in the summer. Sites with less forest tend to be more sensitive in the summer months.

Chapter 4

Missing Data Imputation using Spatial-Temporal Varying Coefficient Model

4.1 Introduction

Water temperature is a determining factor in water quality and may be one of the most important inputs in modeling the impact of climate change on hydrologic systems (Keleher and Rahel, 1996; Beitinger *et al.*, 2000; Flebbe *et al.*, 2006; Meisner, 1990; Caissie, 2006; Minns *et al.*, 1995; Mohseni *et al.*, 1998; Sinokrot and Stefan, 1993). An efficient way to study water temperature is through a real-time monitoring system (Wang *et al.*, 2013). One common method of monitoring water temperature is using modern sensors (Dunham *et al.*, 2005; Huff *et al.*, 2005). However, due to failure of equipment, missing values in water temperature widely occur in practice. Missing data in water temperature can cause problems in studying the hydrologic system and the effects of water temperature on temperature

sensitive aquatic species. Brook trout, for example, prefers cooler water found in high elevation streams, and temperatures greater than $21°C$ are viewed as highly stressful to the health of trout (Meisner, 1990; Beitinger *et al.*, 2000). If missing values occur in the summer period, it limits the ability to make strong inference about the survival of trout. Missing data imputation is thus important. However, infilling missing values in water temperature can be difficult, especially when there are missing values for large gaps in time.

There are two commonly used methods for infilling missing water temperature: taking advantage of the spatial-temporal correlation of water temperature at multiple sites and using other factors to predict water temperature. For the first method, the water temperature is measured at multiple locations and time points thus is spatially and temporally correlated. Water temperatures from neighbor sites or time points could be used to infill missing values. The variables that may be helpful in modeling spatial correlation include latitude and longitude. For the second method, air temperature often has complete data records and is strongly correlated with water temperature. Therefore, it can be used to infill water temperature with reasonable accuracy (Mohseni *et al.*, 1998; Webb *et al.*, 2003). Although there are many studies on missing value imputation for bivariate spatial-temporal data, few of them take advantage of the water-air relationship and the strong autocorrelation of the water temperature simultaneously.

Statistical models can be useful in describing the spatial-temporal correlation of water temperature and water-air relationship, thus are popular for missing values imputation. Those models include simple methods such as the linear regression model and the advanced models such as neural networks. To develop a proper statistical model for infilling the missing data, it is important to take the consideration for the strength of the spatial-temporal correlation of water temperature, correlation between water-air temperature, the distribution of the data and the computational cost.

Some classical statistical models are easily implemented and often can serve as a baseline for performance in the missing data problem. A linear regression model using air temperature as the regressor is easy to interpret and sometimes can obtain moderate prediction accuracy (Neumann *et al.*, 2003). However, such an approach ignores the space and time effects and may not be appropriate for water temperature that is strongly autocorrelated (Li *et al.*, 2014). Another simple model considering temporal correlation is the ARMA model (Cressie and Wikle, 2011). This time series model takes temporal correlation into account but often fails to incorporate the water-air relationship into the model (Li *et al.*, 2014).

Models based on spatial correlation (for example, the Gaussian Process or Kriging (Cressie, 1993)) are widely applied in missing data imputation (Cressie, 1993; Cressie and Wikle, 2011). The Gaussian Process (GP) captures both the linear water-air relationship in the mean component and the nonlinear spatial-temporal correlation of water temperature in the covariance component. Therefore it could achieve accurate results in the missing data imputation. However, it is often challenging to estimate the covariance matrix for large data sets due to the computational difficulty, especially when the GP is applied to bivariate or multivariate time series data (Kaufman *et al.*, 2008).

The neural network (NN) is another popular method and was applied recently to spatial-temporal environmental data (Coulibaly and Evora, 2007). Neural networks are flexible and capable of modeling both the water-air relationship and the spatial-temporal correlation. Also, as the NN does not need a complete physical understanding of those relationships (Diamantopoulou *et al.*, 2007); it works as a black-box for the inputs and outputs. A overview of different NN models and an application to missing data imputation can be found in Coulibaly and Evora (2007).

The varying coefficient model (VCM) (Hastie and Tibshirani, 1993; Fan and Zhang, 2008) is a nonparametric approach to deal with spatial-temporal data. It usually uses a linear

form between a dependent and independent variable, and was shown to be appropriate to describe the relationship between water and air temperature (Li *et al.*, 2014). Moreover, the nonparametric form of the coefficients in the VCM treats space and time covariates as inputs, thus can model spatial and temporal correlation properly (Lu *et al.*, 2009; Serban, 2011). The fitting of the model is straight-forward and computationally fast (Fan and Zhang, 2008; Li *et al.*, 2014).

In this chapter, I propose to use the spatial-temporal VCM (STVCM) to infill missing values in water temperature. A good feature of this method is that the fitting process does not require a complete training data set. Correlation between data from multiple sensors is determined by the time step and distance between sensors. Because of the daily time step, water temperature at the daily scale, usually has strong temporal correlation but weaker spatial correlation due to larger spatial distance. I choose a polynomial spline to model the time effect and kernel smoothing to fit the space effect in the varying coefficients (Hastie *et al.*, 2009). The polynomial splines with parsimonious expression could easily capture the temporal correlation in water temperature (Li *et al.*, 2014). The kernel method for the space effect serves as a tool to borrow strength from neighbor sites to infill missing values. Specifically, for each site with missing water temperature, I choose the nearest neighbor site with complete data to complement the missing part in the target site. The bandwidth is selected based on the adaptive nearest neighbor bandwidth to include the site that is most relevant to the target site (Fan and Gijbels, 1996). This adaptive, self-learning fitting algorithm makes missing data imputation results accurate.

4.2 Data

In this chapter, I evaluate the proposed method by using a simulation study and a real data study. In the simulation study, 35 streams with complete records (for both water and air temperature) are used; their locations are shown in Figure 4.1. For each site, I have a full year of data with the same starting date (1 January 2011) and ending date (31 December 2011). This data will be used to create streams with missing values and evaluate the effect of removal on estimation. In the real data example, I infill the missing water temperature for all our 156 streams. The data used is also a full year record for the year 2011. In both of studies, for each stream, I have $T = 365$ paired daily maximum air and water temperatures for each site.

4.3 Method

4.3.1 Spatial-Temporal Varying Coefficient Model

Let $W_{s,t}$ be the maximum water temperature and $A_{s,t}$ be the maximum air temperature for site s at time t, $t = 1, 2, \ldots, T$, $s = 1, 2, \ldots, S$, and T is the number of time points and S is the number of sites. I consider the following spatial-temporal varying coefficient model for the air-water temperature relationship as

$$W_{s,t} = \theta_0(s,t) + A_{s,t}\theta_1(s,t) + \epsilon_{s,t}, \tag{4.1}$$

where $\theta_0(s,t)$ and $\theta_1(s,t)$ are varying intercept and slope coefficients and $\epsilon_{s,t}$ is the error term in the model. I assume $E(\epsilon_{s,t}) = 0$ and $\text{var}(\epsilon_{s,t}) = \sigma^2$.

For the maximum water temperature, I studied the autocorrelation function and covariance

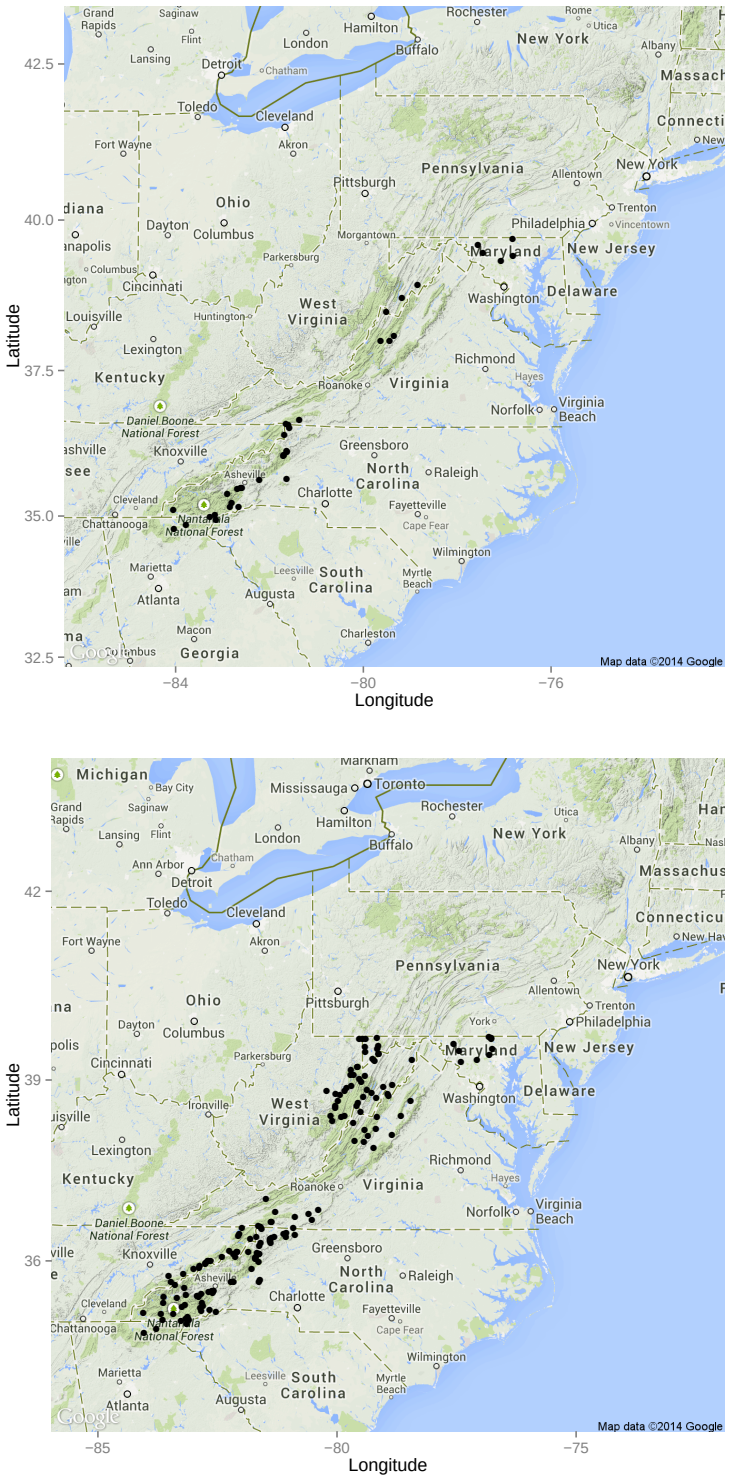

Figure 4.1: Location of 35 streams with complete records for simulation and 156 streams for real data study (black dots are the locations).

matrix among sites. I observed that correlation is strong for time and weak for space. To capture most of the variation in maximum water temperature, the time effect needs to be emphasized. Therefore, I consider the varying coefficients as

$$\theta_0(s,t) = \sum_{j=1}^{K} \alpha_j(s) b_j(t), \theta_1(s,t) = \sum_{j=1}^{K} \beta_j(s) b_j(t). \tag{4.2}$$

Here I use $\{b_1(t), \cdots, b_K(t)\}$ as a set of K basis functions for the time effect to emphasize the time effect. And I use $\alpha_j(s), \beta_j(s)$, $j = 1, 2, \cdots, K$, as coefficients for the basis functions which are varying with space index s. In this sense, I consider the spatial effect as a varying coefficient of the bases for the time effect. Therefore, equation (4.1) can be treated as a linear relationship between water temperature and the covariates

$$\{b_1(t), \cdots, b_K(t), A_{s,t} b_1(t), \cdots, A_{s,t} b_K(t)\}. \tag{4.3}$$

The space coefficients

$$\{\alpha_1(s), \cdots, \alpha_K(s), \beta_1(s), \cdots, \beta_K(s)\} \tag{4.4}$$

are the varying coefficients for the covariate set (4.3).

For the time coefficients (4.3), I use a regression spline for estimation. Specifically, I adopt the quadratic spline method in Li *et al.* (2014) and use the following basis functions:

$$\{b_1(t), \cdots, b_K(t)\} = \{1, t, t^2, (t-\xi_1)_+^2, \cdots, (t-\xi_N)_+^2\}, \tag{4.5}$$

where, $\xi_1, \xi_2, \cdots, \xi_N$ are N knots and $(t-\xi_n)_+, n = 1, 2, \cdots, N$ are the splines with $(t-\xi_n)_+ = t - \xi_n$ if $t \geq \xi_k$ and $(t-\xi_n)_+ = 0$ if $t < \xi_n$.

To select the number of knots N, I adopt the method in Li *et al.* (2014) and fix $N = 4$.

The location of the knots are evenly distributed in time. As part of a sensitivity analysis I compared fits using $N = 3$ and $N = 5$ knots as well as different degrees of polynomials for the spline models and chose to use $N = 4$ knots with quadratic splines as these choices resulted in smooth curves, parsimony and good cross-validation statistics.

For the space coefficients (4.4), I use the local kernel method for estimation (Fan and Zhang, 2008). The reason of not considering a spline basis for (4.4) is the spatial correlation of water temperature is not strong. A spline basis for the space effect would make the shape of the fitted varying coefficients unreliable. It will highly be dependent on the chosen bases. Here I use the local constant method and express, for a fixed u:

$$\alpha_j(u) = \alpha_{j,u}, \beta_j(u) = \beta_{j,u}, u = 1, 2, \cdots, K. \tag{4.6}$$

As I model the correlation of water temperature mainly by temporal correlation, the information used from neighboring sites should be similar to the target site. Therefore, for the bandwidth h, I adopt the adaptive nearest neighbor bandwidth in Fan and Gijbels (1996). For site u, find all the sites that in the neighborhood (the radius is r). Denote the number of sites in the neighborhood by M, with $M > k$. For each time point i, $i = 1, 2, \cdots, T$, and for each of the k sites group in the neighborhood (there are C_M^k combinations), denote by n_i, the number of complete records (note there is at least one non-missing point for the k sites) at time i. Use the combinations for k sites with $\min n_i > 0$. If there are multiple combinations satisfying the condition, choose the nearest one. If there is no combination satisfying the condition, use the k sites that minimize the number of zero $n_i's$. Denote the selected k sites set by B_k and let

$$h = \max_{s \in B_k} \|u - s\|, \tag{4.7}$$

where $\|u - s\|$ is the Euclidean distance between site s and site u.

Then, for a fixed u, the varying coefficients can be estimated by minimizing

$$L(\alpha, \beta) = \sum_t \sum_s [W_{s,t} - \sum_{j=1}^{K} \alpha_j b_j(t) - \sum_{j=1}^{K} \beta_j b_j(t) A_{s,t}]^2 K_h(|u - s|) I_{s \in B_k} + \lambda(\|\alpha\|_2^2 + \|\beta\|_2^2). \tag{4.8}$$

Here $\alpha = (\alpha_{1,u}, \cdots, \alpha_{K,u})$, $\beta = (\beta_{1,u}, \cdots, \beta_{K,u})$, $K_h(x)$ is the kernel function with bandwidth h (I use Epanechnikov kernel $K_h(x) = \frac{3}{4}(1 - x^2)_+ I_{[|x|<h]}$ in this study), λ is a smoothing parameter, $\|\cdot\|_2$ is L_2 norm and I is an indicator function.

The choice of the local constant method instead of local linear or other complex expression is under the consideration of pursuing a parsimonious model. The data from other sites does not help to predict water temperature but to help fit the time effect. In other words, I want to choose the bandwidth h as small as possible such that I can select k sites from the neighborhood to borrow information. The algorithm for estimating the parameters in the model has the following steps:

Algorithm 4.1

Step 1, Determine r and k.

Here r is the radius for neighborhood and k is the number of neighboring sites I select.

Step 2, Calculate the bandwidth in (4.7).

Step 3, Minimize the objective function in (4.8).

Step 4, Repeat step 2, 3 and 4 for all the sites.

4.3.2 Tuning Parameter selection

There are two tuning parameters, k and λ to select before fitting. Due to the weak spatial correlation and strong seasonal pattern of the water temperature, I fix $k = 1$. It is actually a "1-nearest-neighborhood" method. A large k may result in inaccurate imputation because a large number of sites is not helpful in fitting the varying coefficients. For the target site with missing data, a large k will increase the bandwidth h and a large h will reduce the weight of the target site in the fitting.

To select λ, I use generalized cross-validation (GCV) (Wahba, 1990) method. I define $GCV(\lambda)$ as

$$GCV(\lambda) = \sum_{s=1}^{S} (\hat{\mathbf{W}}_s - \mathbf{W}_s)'(\hat{\mathbf{W}}_s - \mathbf{W}_s)/(1 - \mathrm{tr}(\mathbf{S}_\lambda)/T). \tag{4.9}$$

Then the optimal tuning parameter λ_{GCV} is the one minimizing $GCV(\lambda)$.

4.4 Simulation Results

In this section, I use 35 sites with complete records and deliberately remove part of the water temperatures and treat them as missing values. The root mean squared errors (RMSE) statistic is used as a criteria for the imputation performance. The RMSE is defined as

$$RMSE = \sqrt{\frac{1}{|\mathcal{Y}|} \sum_{i \in \mathcal{Y}} (W_i - \hat{W}_i)^2}, \tag{4.10}$$

where $\mathcal{Y}$ is the test set and $|\mathcal{Y}|$ is the number of the observations in set $\mathcal{Y}$ (i.e., the number of missing water values). $W_i's$ are the water temperatures I removed from the original data set (and I did not use them in the training) and $\hat{W}_i's$ are the infilled values.

The models under comparison for missing data imputation are the linear regression model

(Neumann *et al.*, 2003), the nonlinear logistic model (Mohseni *et al.*, 1998), the SAS MI procedure (Allison, 2005), the Gaussian process (Cressie, 1993), the neural network (NN) (Hastie *et al.*, 2009) and the proposed STVCM. For the linear regression model and the nonlinear logistic model, I first fit the model using the complete pairs of data and then infill the missing water by air temperature and the fitted model. For the Gaussian process, I use the linear form of the air temperature for the mean part and Gaussian kernel function for the covariance part. I assume that the temporal correlations of the water temperature series for each site are the same for ease of computing. For the NN, I used the Matlab Neural Network Toolbox (The MathWorks, Inc., 2014) under the default setting. For the STVCM, I used the Algorithm 4.1. I compare the performance of the models under three scenarios which are shown in Table 4.1. These three scenarios are for different types of missing for each site and within each scenario there are three different portions of missing. I denote the three scenarios as S1, S2 and S3.

4.4.1 A Pilot Study

Most of the sites in our study have missing values either at the beginning or at the end of the study period. In this pilot study, for each site, I randomly remove 30% of the water temperature values from either the beginning or the end of the water series and calculate the RMSE for each site. The simulation has only one replicate for each site. The prediction performance for different models is shown in Figure 4.2. From Figure 4.2, I find that the proposed STVCM has the lowest RMSEs and the NN is the only one comparable to STVCM. Therefore, I will compare the performance of STVCM and NN in a more comprehensive simulation design as follows.

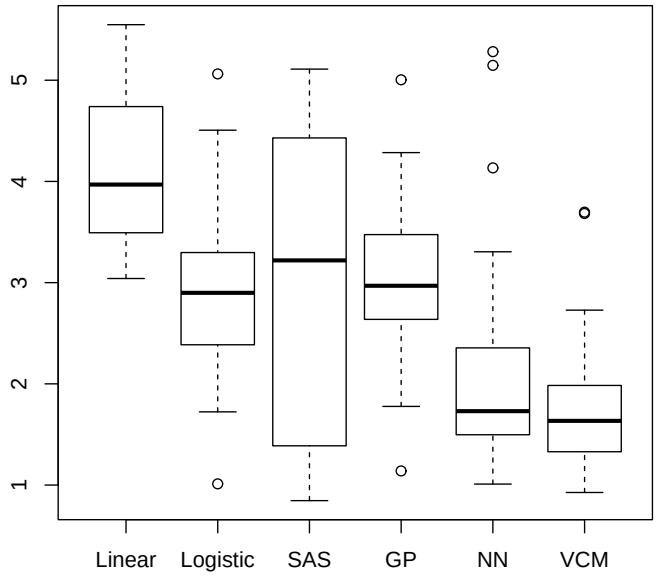

Figure 4.2: Boxplots for RMSEs for six different models (the lower the better).

4.4.2 STVCM vs NN

In this section, I evaluate the performance for STVCM and NN. For each scenario in Table 4.1, I generate the random position of the missing values, calculate the RMSE and repeat for 1000 iterations. The computation time for both methods is shown in Table 4.2 and the RMSEs are shown in Figure 4.3.

Figure 4.3 shows the RMSEs (lower is better) for both the STVCM and the NN. For S1 in Figure 4.3, I find that the imputation performance of the STVCM is better than the NN. The STVCM in equation (4.2) uses a smooth method in the varying intercept $\theta_0(s,t)$ that is able to capture the main trend of the water temperature. As the missing values at the random locations do not affect modeling (smoothing) the trend of water temperature, the STVCM thus takes advantage of the highly correlated water temperature and easily borrows information from the neighboring time points to predict the missing values. Therefore it is not surprising that the STVCM has accurate imputation results. For S2 in Figure 4.3, I find that the NN is slightly better than the STVCM in the imputation. Due to the nature of the regression spline used in equation (4.5), one big gap of missing values will affect the smoothing of the trend for water temperature. Therefore, it is difficult for both models to borrow information from neighboring time points. Both models have to rely on the air-water relationship for prediction. Also I can see that in part (d), (e) and (f), the performance of the proposed method degenerates as the percent of missing increases. For S3 in Figure 4.3, I find that the STVCM achieves better results than the NN. In this scenario, the missing water temperature is outside of the time range of the non-missing data. The algorithm used in the STVCM considers the spatial correlation between sites and searches for the best neighbor to use to infill the missing parts. It works more efficiently and precisely to select the most useful information from the neighborhood. In contrast, the NN still uses all the available information and may lead to bad performance in some cases which can be seen

from the outliers in 4.3, part (i). Overall, the STVCM is more appropriate except for one or two extreme missing values scenarios. For more justification of this conclusion and to help visualize the difference, I define the paired RMSE by subtracting the NN RMSE from the STVCM RMSE for each simulation. The paired RMSE plot is in Figure 4.4.

Another advantage of the STVCM is that it is much faster than the NN in terms of computational time. Table 4.2 shows the simulation time for all the nine scenarios. Because the STVCM is relatively simple, the computational time is about $80\% - 90\%$ lower than the NN.

Table 4.1: Number and type of missing for each site for the simulation study

Missing location \ Percent missing	10%	20%	30%
Randomly (S1)	10% at random	20% at random	30% at random
One gap at random location (S2)	10% one gap	20% one gap	30% one gap
One gap at beginning or end (S3)	10% at beginning or end	20% at beginning or end	30% at beginning or end

Table 4.2: Computation time for STVCM and NN methods (in seconds).

	10%		20%		30%	
	STVCM	NN	STVCM	NN	STVCM	NN
Randomly (S1)	6579	39461	6519	35536	6439	30700
One gap at random location (S2)	4525	46276	4521	45282	4483	30306
One gap at beginning or end (S3)	5242	39899	5183	37254	5137	31666

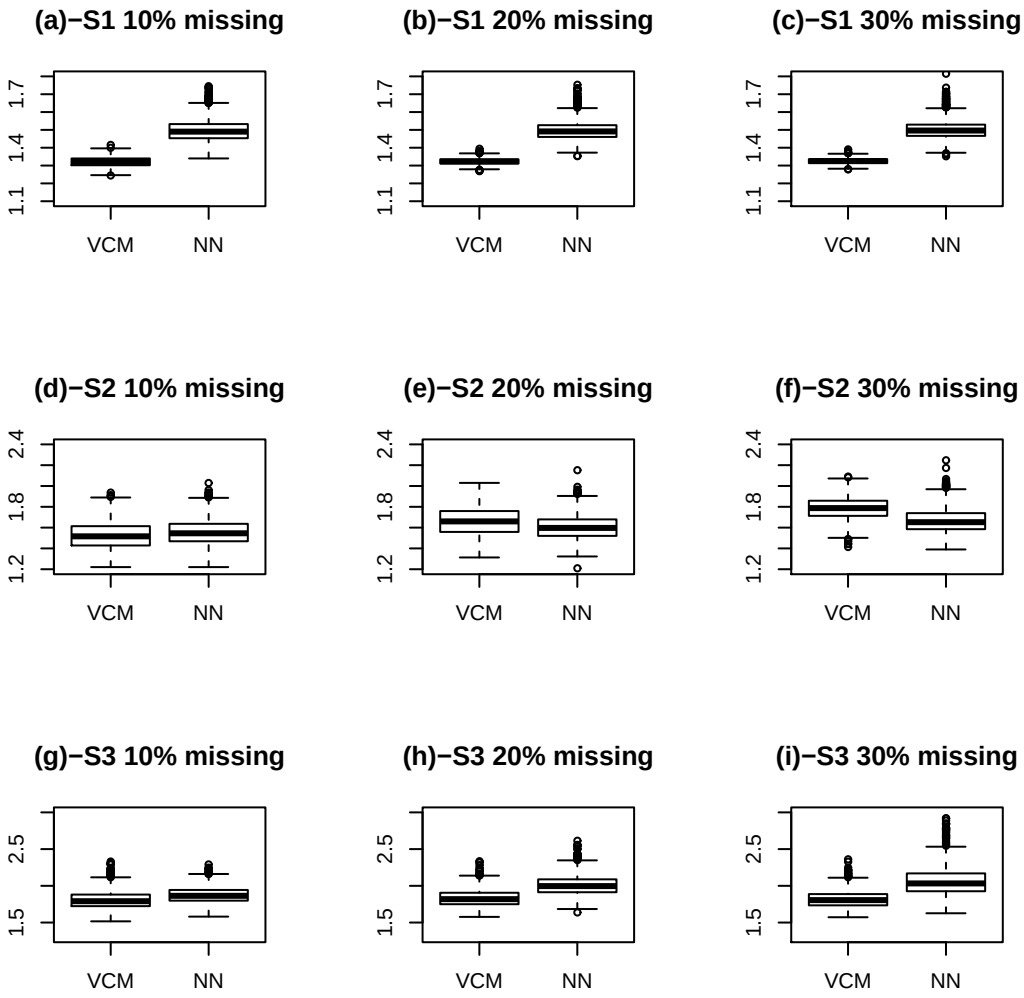

Figure 4.3: Boxplots for RMSEs for STVCM and NN. (a) 10 percent missing at random locations. (b) 20 percent missing at random locations. (c) 30 percent missing at random locations. (d) 10 percent missing in one gap. (e) 20 percent missing in one gap. (f) 30 percent missing in one gap. (g) 10 percent missing at the beginning or the end of the time series. (h) 20 percent missing at the beginning or the end of the time series. (i) 30 percent missing at the beginning or the end of the time series.

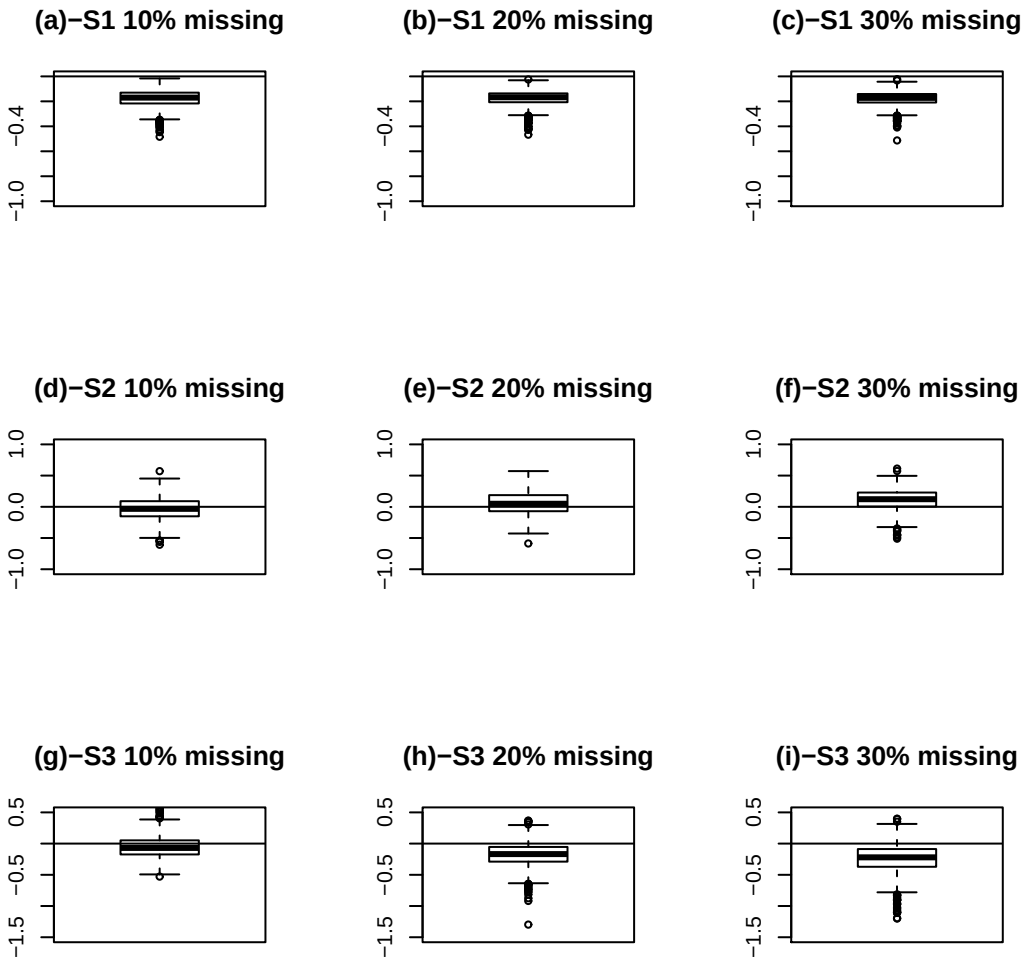

Figure 4.4: Boxplots for Paired RMSEs for STVCM and NN (negative values mean STVCM performs better). (a) 10 percent missing at random locations. (b) 20 percent missing at random locations. (c) 30 percent missing at random locations. (d) 10 percent missing in one gap. (e) 20 percent missing in one gap. (f) 30 percent missing in one gap. (g) 10 percent missing at the beginning or the end of the time series. (h) 20 percent missing at the beginning or the end of the time series. (i) 30 percent missing at the beginning or the end of the time series.

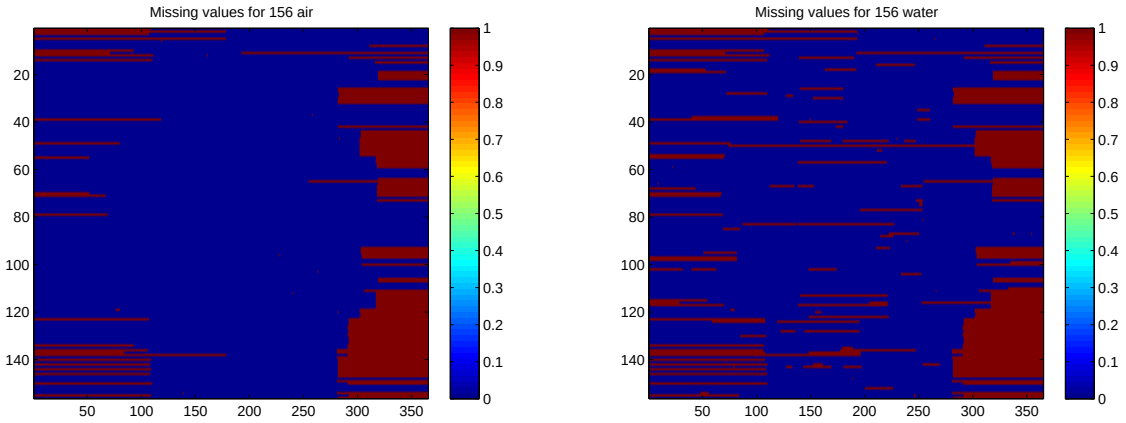

Figure 4.5: Heat map for missing data for the 156 sites. The y-axis is the site numbers and the x-axis is the day.

4.5 Real Data Study

In this section, I applied the clustering methods developed in chapter 3 and the proposed STVCM to all our 156 streams. For all those streams, the percent of missing data for the air and water temperature data are shown in the heat map in Figure 4.5. I can see that most of the streams have both air and water temperature missing from October to December, 2011. So I extract nine months of data starting from January 1st to September 30th. The total number of days in this study is 273. I use the SAS MI procedure to obtain the complete air temperature. The missing water temperature will be infilled by our spatial-temporal STVCM.

Figure 4.6 shows the intercept and slope curves for all of the 156 sites obtained by spatial-temporal VCM. The pattern of both curves are similar to the those in Figure 3.2, which is strong evidence that the 1-nearest neighborhood methods in the parameter estimation produces the varying coefficients close to that in the STVCM for single sites. The profile shown in Figure 4.6 is the data used for the following clustering analysis. I then apply the method introduced in section 3.3.2 to the intercept and slope curves (bivariate curve). Figure

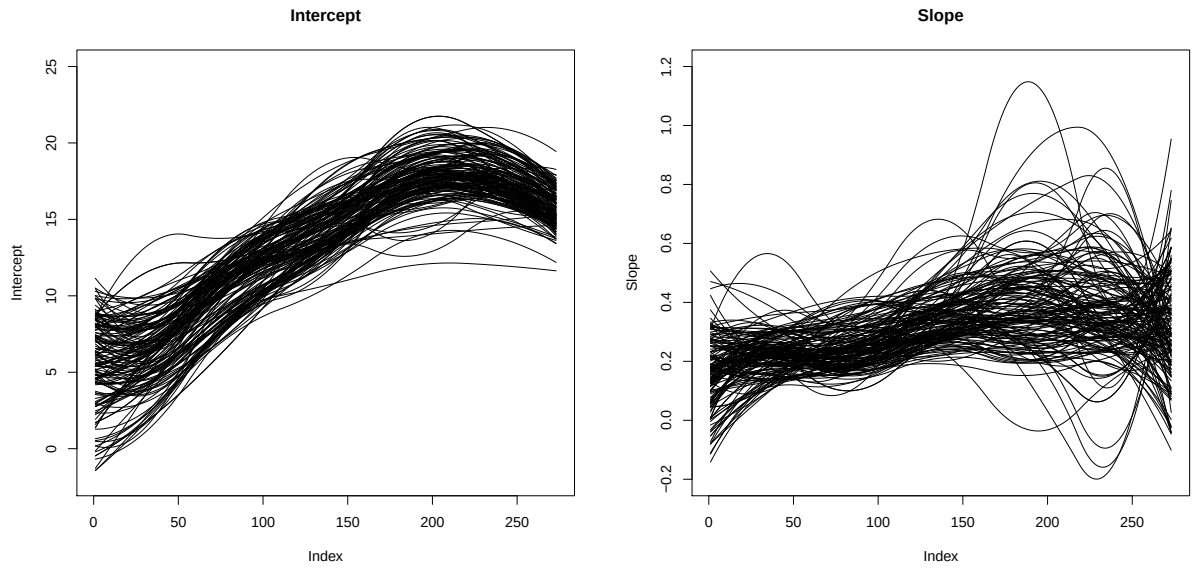

Figure 4.6: Estimated varying coefficients for the 156 streams

4.7 shows the estimated variogram. For the variogram under $w = 0$, there is a flat pattern in the estimated values. That means the slope curves do have very weak spatial correlation.

Clustering Result

The K-Medoids clustering method is applied on the distance measure (3.2). The silhouette statistics for choosing the number of clusters are summarized in Figure 4.8. Figures 4.9-4.13 report the performance of the proposed clustering method in chapter 3 with respect to the weight $w = 100\%, 75\%, 50\%, 25\%$ and 0% on the intercept. Recall that in chapter 3 I have three major findings on the clustering results. They are discussed in this section.

First, the weight can determine the impact of spatial correlation on the cluster results. This statement is generally true for all our 156 streams except for several streams in the north. Second, the intercept and slope curves have connections with the landscape and climate variables in the clustering results. It can be verified through Table 4.3, which is similar to Table 3.1. Third, the distance measure (3.2) has more advantages if I place considerable

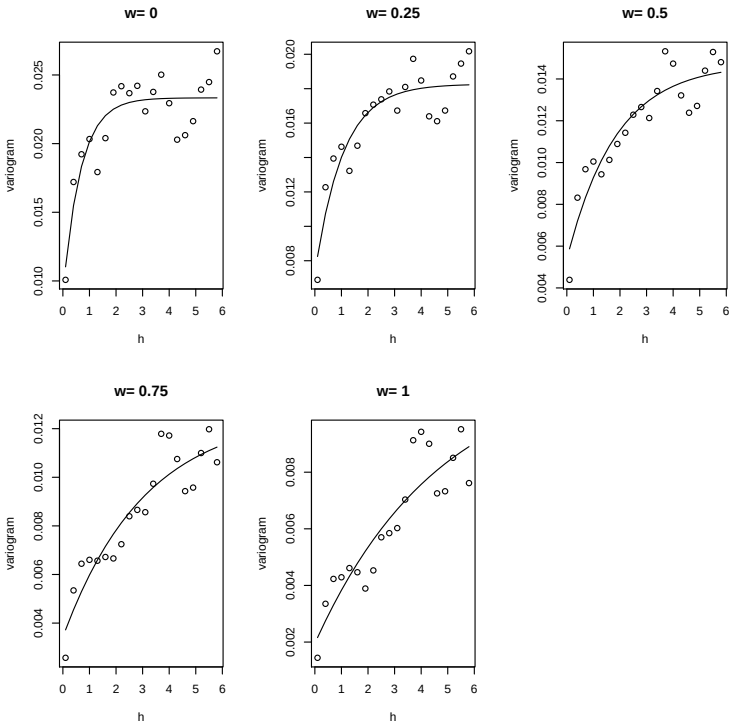

Figure 4.7: Variogram estimates and parametric form for different weights based on the intercept curve: circles are empirical estimates of the variogram and curves are for the estimated exponential variogram.

amount of weight, i.e., $w = 50\%$, on both intercept and slope. From the boxplots in part (d), (e) and (f) of Figure 4.12 and Table 4.3, elevation, solar radiation and percent forest exhibit significant differences across clusters. Therefore, the distance measure (3.2) can also provide meaningful clustering results for all 156 sites. Another interesting finding when $w = 50\%$ is that there is a fourth cluster.

In Figure 4.12, part (c), I can see that the fourth cluster is a top layer, which means high sensitivity. That would explain the high boxplot in part (d) and low boxplot in part (e) for the fourth cluster. However, the ranges for the boxplots for the fourth cluster is huge in part (d), (e) and (f), almost from the minimum to the maximum. Because the silhouette statistics select $k = 4$, the fourth cluster might be related to some other landscape or climate variables which I have not considered in this study. As a result, the fourth cluster needs to be further investigated.

Table 4.3: F statistics and p-values (in parenthesis) from F-test for cluster differences

Weight on intercept	Solar radiation	Percent forest	Elevation
0	20.9(<0.01)	25.1(<0.01)	0.05(0.81)
0.25	18.2(<0.01)	20.7(<0.01)	1.38(0.25)
0.5	13.0(<0.01)	21.0(<0.01)	2.84(0.03)
0.75	4.45(0.03)	0.20(0.64)	22.2(<0.01)
1	1.59(0.20)	1.81(0.18)	13.3(<0.01)

4.6 Conclusion

In this chapter, I developed a spatial-temporal varying coefficient model (STVCM) to infill the missing values in the water temperature and apply it to all the 156 streams in this study to obtain a complete data record. The STVCM models both the temporal and the spatial

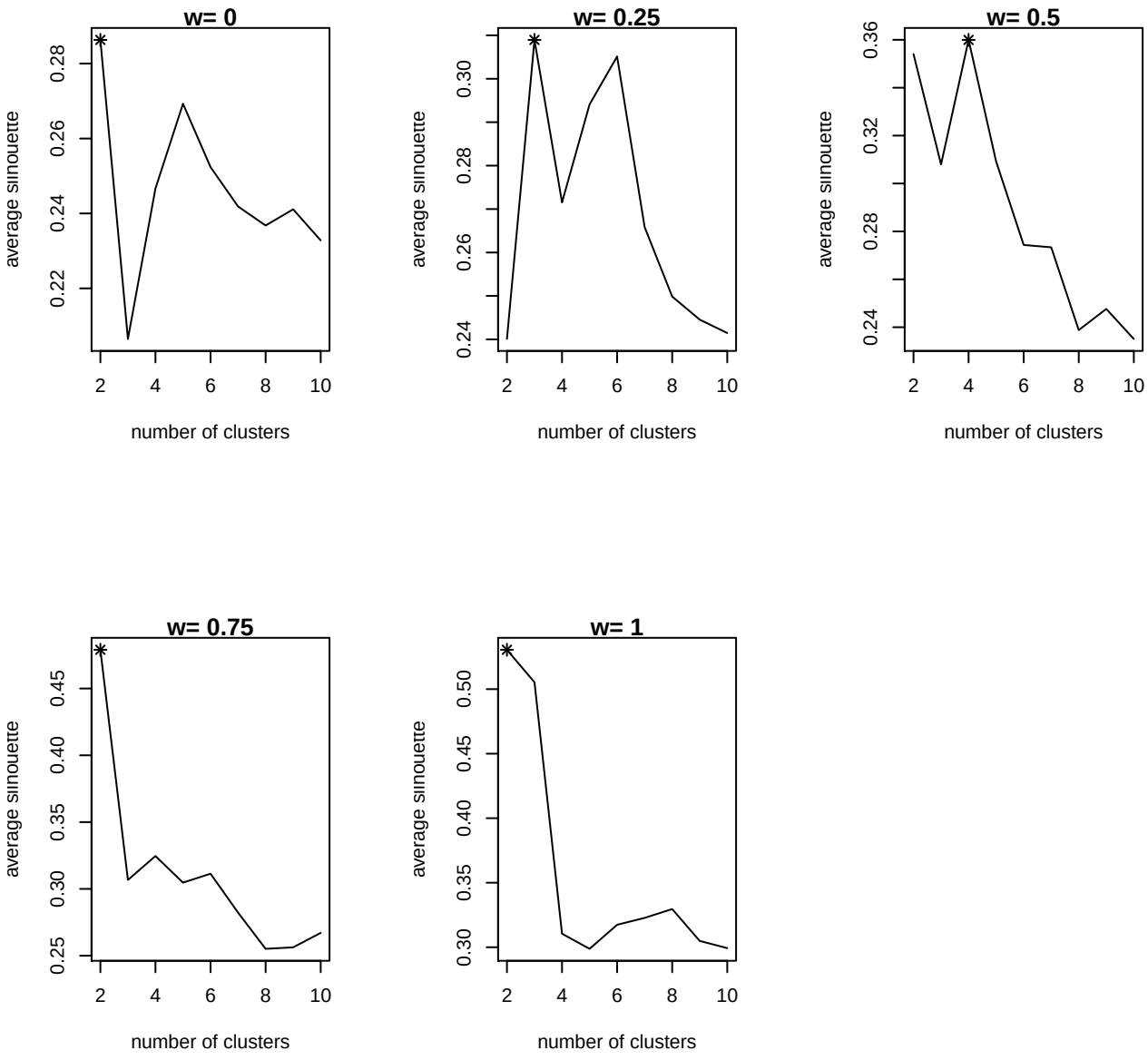

Figure 4.8: Silhouette statistics for different numbers of clusters under five different weights. The star marks indicate the optimal number of clusters and the corresponding Silhouette statistics.

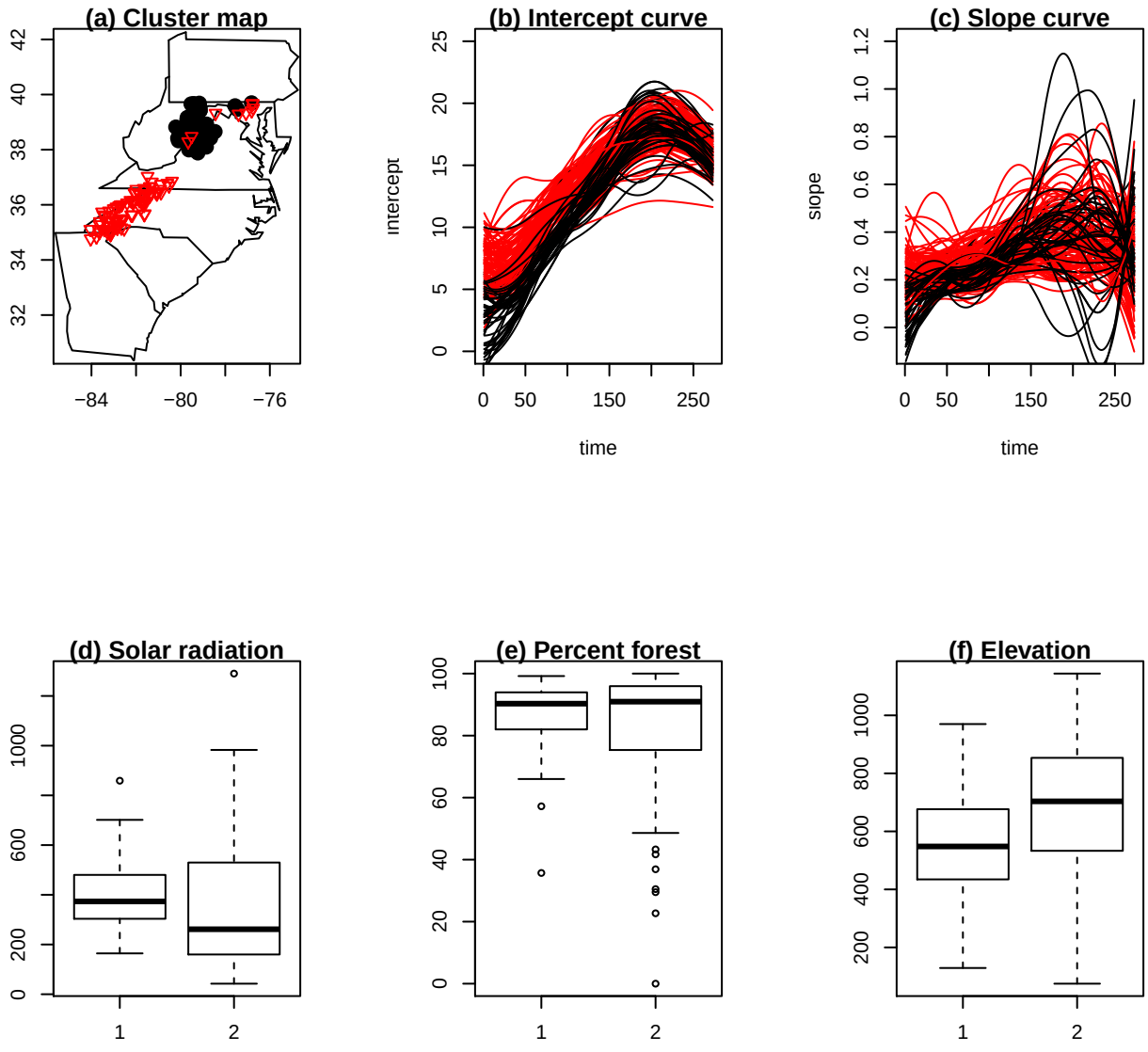

Figure 4.9: Cluster results for weight=100% on intercept. Numbers of sites in each cluster are: 56 for cluster 1 (black) and 100 for cluster 2 (red). (a) location of streams. (b) intercept curves. (c) slope curves. (d) boxplots for solar radiation for different clusters. (e) boxplots for percent forest for different clusters. (f) boxplots for elevation for different clusters.

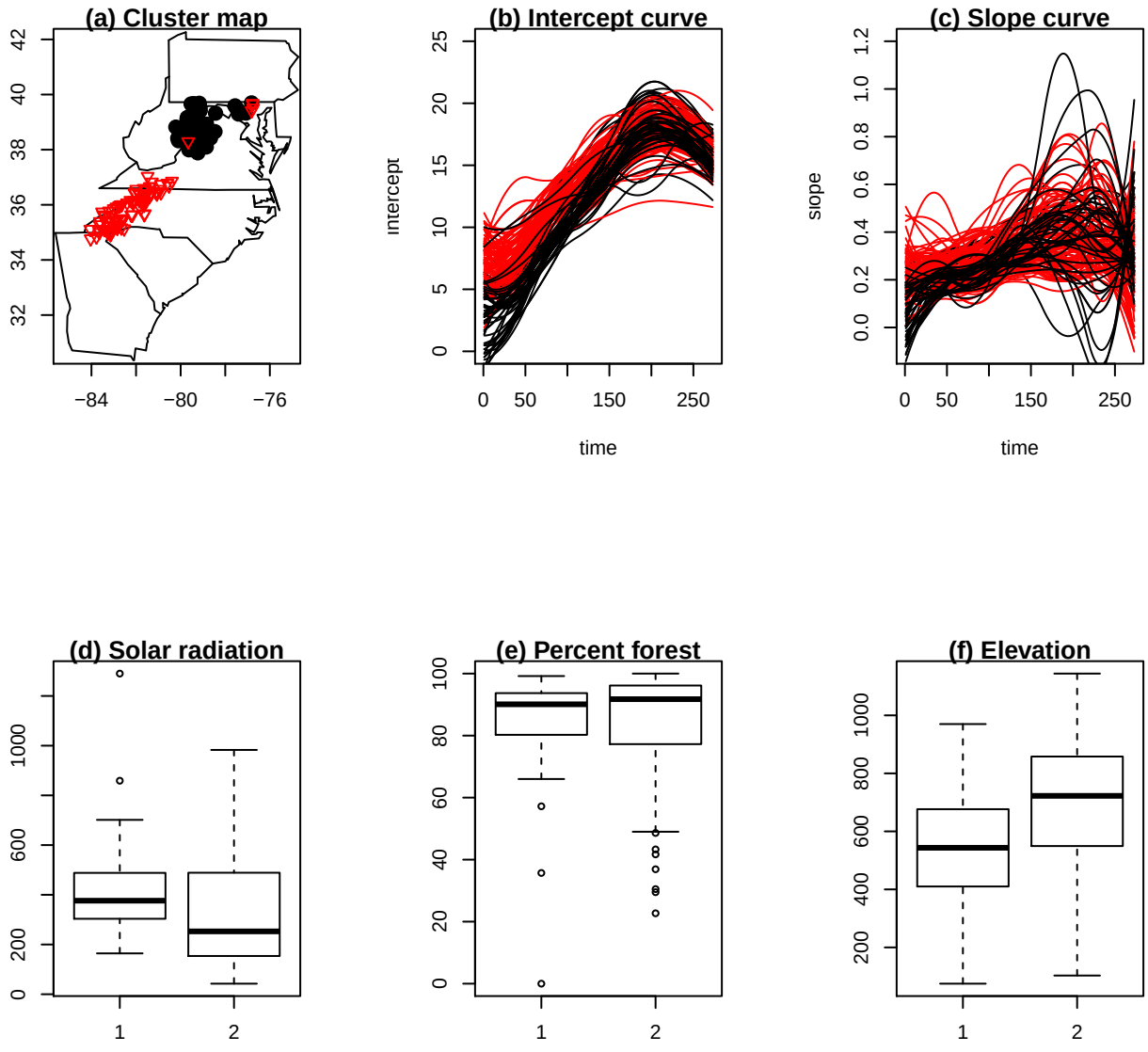

Figure 4.10: Cluster results for weight=75% on intercept. Numbers of sites in each cluster are: 60 for cluster 1 (black) and 96 for cluster 2 (red). (a) location of streams. (b) intercept curves. (c) slope curves. (d) boxplots for solar radiation for different clusters. (e) boxplots for percent forest for different clusters. (f) boxplots for elevation for different clusters.

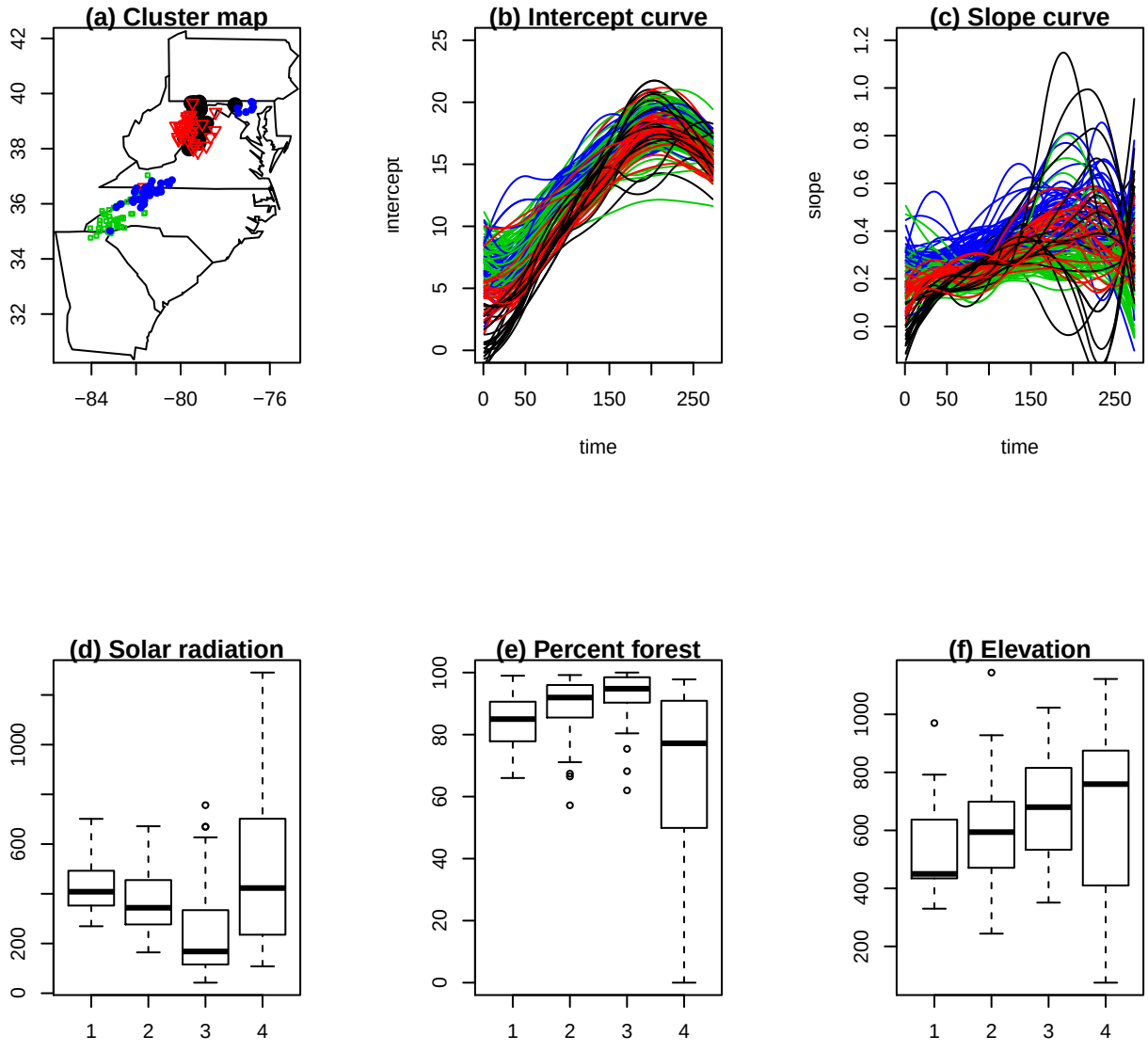

Figure 4.11: Cluster results for weight=50% on intercept. Numbers of sites in each cluster are: 24 for cluster 1 (black), 34 for cluster 2 (red), 52 for cluster 3 (green) and 46 for cluster 4 (blue). (a) location of streams. (b) intercept curves. (c) slope curves. (d) boxplots for solar radiation for different clusters. (e) boxplots for percent forest for different clusters. (f) boxplots for elevation for different clusters.

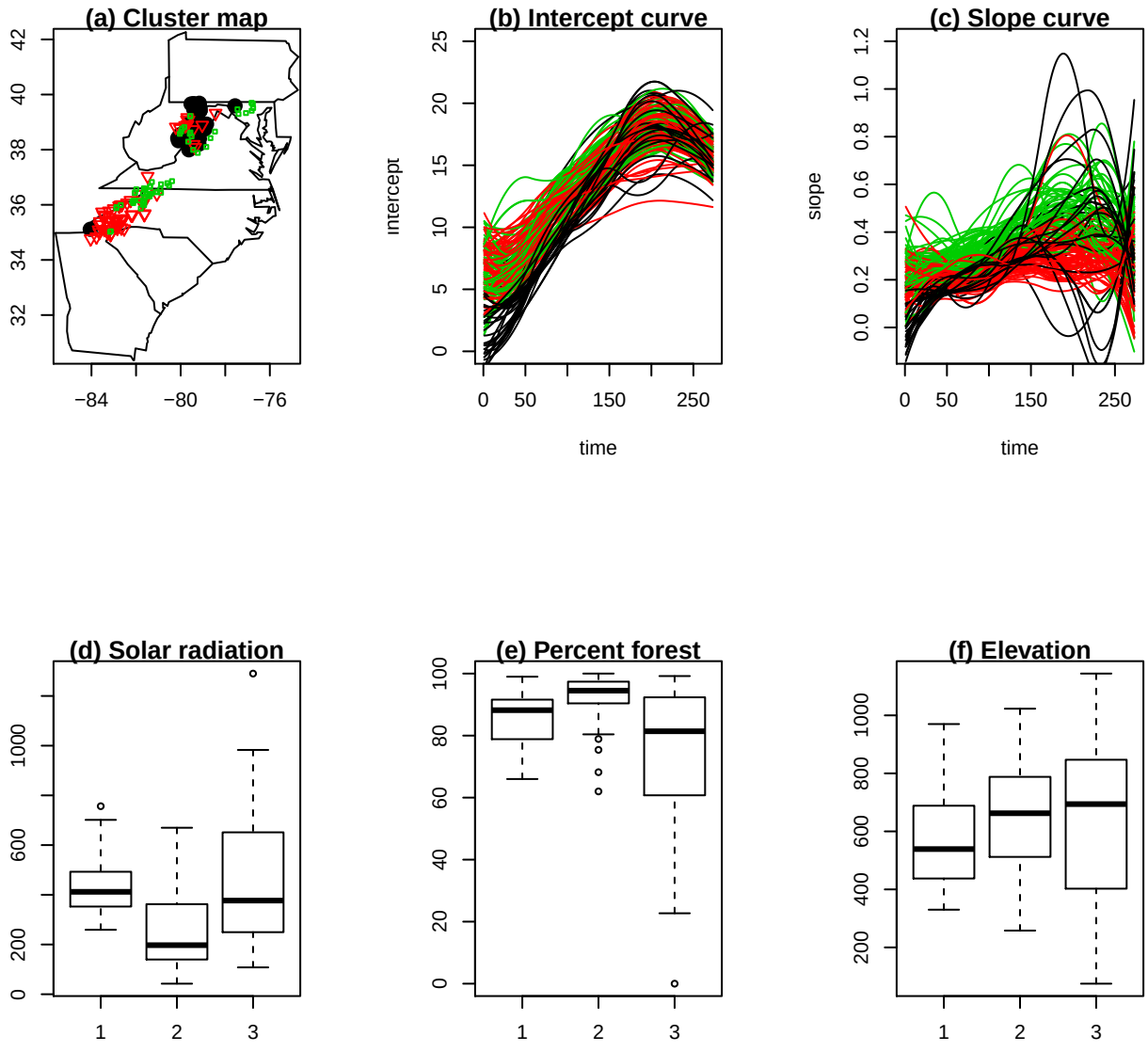

Figure 4.12: Cluster results for weight=25% on intercept. Numbers of sites in each cluster are: 31 for cluster 1 (black), 62 for cluster 2 (red) and 63 for cluster 3 (green). (a) location of streams. (b) intercept curves. (c) slope curves. (d) boxplots for solar radiation for different clusters. (e) boxplots for percent forest for different clusters. (f) boxplots for elevation for different clusters.

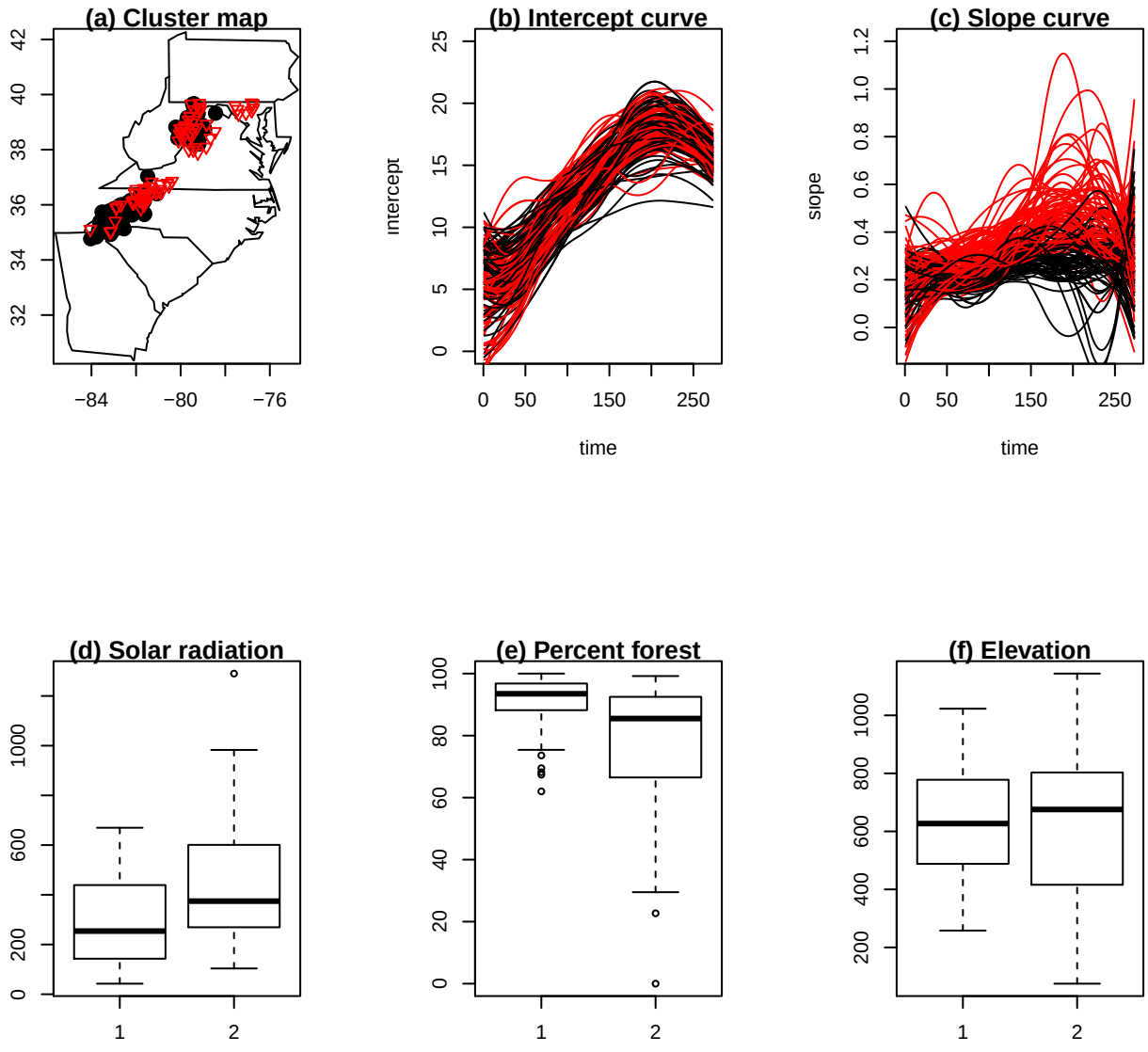

Figure 4.13: Cluster results for weight=0% on intercept. Numbers of sites in each cluster are: 70 for cluster 1 (black) and 86 for cluster 2 (red). (a) location of streams. (b) intercept curves. (c) slope curves. (d) boxplots for solar radiation for different clusters. (e) boxplots for percent forest for different clusters. (f) boxplots for elevation for different clusters.

variation of the water temperature thus provides an accurate approach for missing data imputation. The simulation study shows that the performance of the STVCM on missing data imputation is better than existing methods such as neural network, Gaussian process, etc.

The water temperature data in this study has strong temporal correlation and weak spatial correlation. The choice of polynomial splines for the time effect and kernel smoothing for the space effect in the varying coefficients is thus reasonable. Polynomial splines capture most of the variation in the water temperature and the kernel method for the space effect is only used to provide data from neighbor sites in case of large gaps. Because the choice of kernel functions and bandwidth is very flexible, the space effect can be chosen to be very weak hence it will not affect the shape of the fitted curve. I also conducted some experiments using polynomial splines for both time and space effects. Although the imputation results remain plausible, the shape of the fitted curve does not provide useful information for the further clustering analysis.

Chapter 5

Future Work

There are some potential developments for the methods presented in this dissertation, which are discussed in this chapter. Those include further inference for the VCM, extending the current method to multivariate data and improving the accuracy of the spatial-temporal VCM for missing data imputation.

For a single site, further analysis techniques might be applied to better understand the power of the VCM to accurately model the air-water temperature relationship. First, point-wise or simultaneous confidence bands, constructed based on the estimated varying coefficients, could be used for predicting the range of the water temperature. Second, the selection of number and location of the knots could be more flexible if the number of days in the study varies. In addition, the selection of basis functions could be improved such that each parameter has physical meanings. Third, as more data are currently being collected using paired thermographs throughout southeast USA, I will extend the current VCM to incorporate additional terms in the model.

The weighted distance in chapter 3 can be extended from bivariate to multivariate functional

data by using more than two weights. One application is to study the relationship between water temperature and other covariates using the VCM and generate more than one slope curve. By assigning different weights to multivariate curves in the distance measure, the resultant clusters will reflect the sensitivity of water temperature to different covariates of interest.

For the spatial-temporal VCM in chapter 4, the number of the neighbors k used in the kernel methods is fixed. Some preliminary studies show that the choice of k will significantly affect the performance of the model in missing data imputation. The optimal k would reduce the RMSE by 20% compared to using $k = 1$. The choice of $k = 1$ will give the best result if the target site and nearest site behave similarly. However, as the spatial correlation between sites is not strong, it is very likely that the nearest site gives little information for infilling missing data in the target site. Another interesting problem is to test whether the intercept curve and/or the slope curve are constant functions. For example, constant slope means that the sensitivity of the water temperature to the change of the air temperature remains constant during a specific period of time. That would be the ideal case for the survival of the fish, especially in summer. Therefore, testing and grouping the streams with constant slopes would be beneficial for water management.

Bibliography

Abraham C, Cornillon P, Matzner-Lober E, Molinari N, 2003. Unsupervised curve clustering using b-splines. *Scandinavian Journal of Statistics. Theory and Applications* **30**: 581–595.

Aggarwal CC, Reddy CK, 2013. *Data Clustering: Algorithms and Applications*. CRC Press, Boca Raton, FL.

Ahmadi-Nedushan B, St-Hilaire A, Ouarda TB, Bilodeau L, Robichaud E, Thiemonge N, Bobee B, 2007. Predicting river water temperatures using stochastic models: Case study of the Moisie River (Quebec, Canada). *Hydrological Processes* **21**: 21–34.

Alexander L, 2013. Working group I contribution to the IPCC fifth assessment report climate change 2013: The physical science basis summary for policymakers. Technical report, Intergovernmental Panel on Climate Change, Stockholm, Sweden.

Allison PD, 2005. Imputation of categorical variables with PROC MI. Technical report, SAS Users Group International, 30th Meeting (SUGI 30), Philadelphia, PA .

Beitinger TL, Bennett WA, McCauley RW, 2000. Temperature tolerances of North American freshwater fishes exposed to dynamic changes in temperature. *Environmental Biology of Fishes* **58**: 237–275.

BIBLIOGRAPHY

Ben-Dor A, Shamir R, Yahkini Z, 1999. Clustering gene expression patterns. *Journal of Computational Biology* **6**: 281–297.

Benyahya L, Caissie D, St-Hilaire A, Ouarda TB, Bobee B, 2007. A review of statistical water temperature models. *Canadian Water Resources Journal* **32**: 179–192.

Benyahya L, St-Hilaire A, Ouarda TB, Bobee B, Dumas J, 2008. Comparison of non-parametric and parametric water temperature models on the Nivelle River, France. *Hydrological Sciences Journal* **53**: 640–655.

Berrendero J, Justel A, Svarc M, 2011. Principal components for multivariate functional data. *Computational Statistics and Data Analysis* **55**: 2619–2634.

Brenden T, Wang L, Seelbach P, Clark RJ, Wiley M, Sparks-Jackson B, 2008. A spatially constrained clustering program for river valley segment delineation from GIS digital river networks. *Environmental Modelling & Software* **23**: 638–649.

Caissie D, 2006. The thermal regime of rivers: A review. *Freshwater Biology* **51**: 1389–1406.

Caissie D, El-Jabi N, St-Hilaire A, 1998. Stochastic modelling of water temperatures in a small stream using air to water relations. *Canadian Journal of Civil Engineering* **25**: 250–260.

Chen YD, Carsel RF, McCutcheon SC, Nutter WL, 1998. Stream temperature simulation of forested riparian areas: 1. Watershed-scale model development. *Journal of Environmental Engineering* **124**: 304–315.

Chenard J, Caissie D, 2008. Stream temperature modeling using artificial neural networks: Application on Catamaran Brook, New Brunswick, Canada. *Hydrological Processes* **22**: 3361–3372.

BIBLIOGRAPHY

Cheng MY, Zhang W, Chen LH, 2009. Statistical estimation in generalized multiparameter likelihood models. *Journal of the American Statistical Association* **104**: 1179–1191.

Chiou JM, Li PL, 2007. Functional clustering and identifying substructures of longitudinal data. *Journal of the Royal Statistical Society. Series B. Statistical Methodology* **69**: 679–699.

Cho HY, Lee KH, 2012. Development of an air–water temperature relationship model to predict climate-induced future water temperature in estuaries. *Journal of Environmental Engineering* **138**: 570–577.

Cluis D, 1972. Relationship between stream water temperature and ambient air temperature a simple autoregressive model for mean daily stream water temperature fluctuations. *Nordic Hydrology* **3**: 65–71.

Coulibaly P, Evora N, 2007. Comparison of neural network methods for infilling missing daily weather records. *Journal of Hydrology* **341**: 27–41.

Cressie N, 1993. *Statistics for Spatial Data,*. John Wiley & Sons., New York.

Cressie N, Wikle CK, 2011. *Statistics for Spatio-Temporal Data*. Wiley, Hoboken, NJ.

Diamantopoulou MJ, Antonopoulos VZ, Papamichail DM, 2007. Cascade correlation artificial neural networks for estimating missing monthly values of water quality parameters in rivers. *Water Resources Management* **21**: 649–662.

Dunham J, Gwynne C, Reiman B, Martin D, 2005. Measuring stream temperature with digital data loggers: A user's guide. Technical report, Report RMRS-GTR-150WWW, USDA Forest Service, Fort Collins, CO.

EBTJV, 2006. *Eastern Brook Trout Joint Venture*.

BIBLIOGRAPHY

Fan J, Gijbels I, 1996. *Local Polynomial Modelling and Its Applications*. Chapman and Hall, New York.

Fan J, Zhang W, 2008. Statistical methods with varying coefficient models. *Statistics and Its Interface* **1**: 179–195.

Ferguson CA, Bowman AW, Scott EM, Carvalho L, 2007. Model comparison for a complex ecological system. *Journal of the Royal Statistical Society, Series A* **170**: 691–711.

Ferguson CA, Bowman AW, Scott EM, Carvalho L, 2009. Multivariate varying-coefficient models for an ecological system. *Environmetrics* **20**: 460–476.

Flebbe PA, Roghair LD, Bruggink JL, 2006. Spatial modeling to project southern Appalachian trout distribution in a warmer climate. *Transactions of the American Fisheries Society* **165**: 1371–1382.

Giraldo R, Delicado P, Mateu J, 2012. Hierarchical clustering of spatially correlated functional data. *Statistica Neerlandica* **66**: 403–421.

Hastie T, Tibshirani R, 1993. Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)* **55**: 757–796.

Hastie T, Tibshirani R, Friedman J, 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, NY.

Hoover DR, Rich JA, Wu CO, Yang LP, 1998. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* **85**: 809–822.

Hornik K, 2013. The R FAQ.

Huang JZ, Wu CO, Zhou L, 2002. Varying-coefficient models and basis function apporximations for the analysis of repeated measurements. *Biometrika* **89**: 111–128.

BIBLIOGRAPHY

Huff DD, Hubler SL, Borisenko AN, 2005. Using field data to estimate the realized thermal niche of aquatic vertebrates. *North American Journal of Fisheries Management* **25**: 346–360.

Ieva F, Paganoni A, Pigoli D, Vitelli V, 2013. Multivariate functional clustering for the morphological analysis of electrocardiograph curves. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **62**: 401–418.

Jacques J, Preda C, 2014. Model-based clustering for multivariate functional data. *Computational Statistics and Data Analysis* **71**: 92–106.

Journel AG, Huijbregts CJ, 1978. *Mining Geostatistics*. Academic Press, London.

Kaufman CG, Schervish MJ, Nychka DW, 2008. Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association* **103**: 1545–1555.

Kayano M, Dozono K, Konishi S, 2010. Functional cluster analysis via orthonormalized Gaussian basis expansions and its application. *Journal of Classification* **27**: 211–230.

Keleher CJ, Rahel FJ, 1996. Thermal limits to salmonid distributions in the Rocky Mountain region and potential habitat loss due to global warming: A geographic information systems (GIS) approach. *Transactions of the American Fisheries Society* **125**: 1–13.

Kothandaraman V, 1971. Analysis of water temperature variations in large river. *ASCE Journal of the Sanitary Engineering Division* **97**: 19–31.

Kothandaraman V, 1972. Air water temperature relationship in Illinois river. *Water Resources Bulletin* **8**: 38–45.

Kunsch HR, 1989. The jackknife and the bootstrap for general stationary observations. *Annals of Statistics* **17**: 1217–1241.

BIBLIOGRAPHY

Kutner MH, Nachtsheim CJ, Neter J, Li W, 2004. *Applied Linear Statistical Models*. McGraw-Hill/Irwin, New York.

Lance GN, Williams WT, 1967. Mixed-data classification programs. i. Agglomerative systems. *Aust. Comput. J.* **1**: 15–20.

Le ND, Zidek JV, 2006. *Statistical Analysis of Environmental Space-Time Process*. Springer, New York.

Li H, Deng X, Kim DY, Smith EP, 2014. Modeling maximum daily temperature using a varying coefficient regression model. *Water Resources Research* **50**: 3073–3087.

Lisle TE, 1987. Using "residual depths" to monitor pool depths independently of discharge. Technical report, Note PSW-394, Pacific Southwest Forest and Range Experiment Station, Forest Service, U.S. Department of Agriculture, Berkeley, CA.

Lu Z, Steinskog DJ, Tjstheim D, Yao Q, 2009. Adaptively varying-coefficient spatiotemporal models. *Journal of the Royal Statistical Society. Series B (Methodological)* **71**: 859–880.

Mayer TD, 2012. Controls of summer stream temperature in the Pacific Northwest. *Journal of Hydrology* **475**: 323–335.

Meisner JD, 1990. Effect of climate warming on the southern margins of the native range of brook trout, *Salvelinus Fontinalis*. *Canadian Journal of Fisheries and Aquatic Science* **47**: 1065–1070.

Minns C, Randall R, Chadwick E, Moore J, Green R, 1995. Potential impact of climate change on the habitat and production dynamics of juvenile Atlantic salmon (*Salmo salar*) in eastern Canada. In Beamish R (ed.), *Climate Change and Northern Fish Population*, NRC Research Press, 699–708.

BIBLIOGRAPHY

Mohseni O, Erickson TR, Stefan HG, 1999. Sensitivity of stream temperatures in the United States to air temperatures projected under a global warming scenario. *Water Resources Research* **35**: 3723–3733.

Mohseni O, Stefan HG, 1999. Stream temperature/air temperature relationship: A physical interpretation. *Journal of Hydrology* **218**: 128–141.

Mohseni O, Stefan HG, Erickson TR, 1998. A nonlinear regression model for weekday stream temperatures. *Water Resources Research* **34**: 2685–2692.

Nash JE, Sutcliffe JV, 1970. River flow forecasting through conceptual models, I-A, discussion of principles. *Journal of Hydrology* **10**: 282–290.

Neumann DW, Rajagopalan B, Zagona EA, 2003. Regression model for daily maximum stream temperature. *Journal of Environmental Engineering* **129**: 667–674.

Oliver M, Webster R, 1989. A geostatistical basis for spatial weighting in multivariate classification. *Mathematical Geology* **21**: 15–35.

Onset Computer Corporation, 2009. *HOBO U22 Water Temp Pro v2 users manual. Document number 10366-C.*

Ramsay J, Silverman B, 2005. *Functional Data Analysis*. Springer Series in Statistics, Springer, New York, second edition.

Ray S, Mallick B, 2006. Functional clustering by bayesian wavelet methods. *Journal of the Royal Statistical Society. Series B (Methodological)* **68**: 305–332.

Rencher AC, Christensen WF, 2012. *Methods of Multivariate Analysis*. Wiley, New York.

Rencher AC, Schaalje GB, 2008. *Linear Models in Statistics*. John Wiley & Sons, Inc., Hoboken, NJ.

BIBLIOGRAPHY

Rousseeuw PJ, 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Computational and Applied Mathematics* **20**: 53–65.

Ruppert D, Wand MP, Carroll RJ, 2003. *Semiparametric Regression*. Cambridge University Press, New York.

Sangalli L, Secchi P, Vantini S, Vitelli V, 2010. K-means alignment for curve clustering. *Computational Statistics and Data Analysis* **54**: 1219–1233.

Serban N, 2011. A spacetime varying coefficient model: The equity of service accessibility. *The Annals of Applied Statistics* **5**: 2024–2051.

Sinokrot BA, Stefan HG, 1993. Stream temperature dynamics: Measurements and modeling. *Water Resources Research* **29**: 2299–2312.

St-Hilaire A, Ouarda TB, Bargaoui Z, Daigle A, Bilodeau L, 2012. Daily river water temperature forecast model with a k-nearest neighbour approach. *Hydrological Processes* **26**: 1302–1310.

Stefan H, Fang X, Eaton J, 2001. Simulated fish habitat changes in North American lakes in response to projected climate warming. *Transactions of the American Fisheries Society* **130**: 459–477.

Stefan H, Preud'homme E, 1993. Stream temperature estimation from air temperature. *Water Resources Bulletin* **29**: 27–45.

Tarpey T, Kinateder K, 2003. Clustering functional data. *Journal of Classification* **20**: 93–114.

The MathWorks, Inc, 2014. *Neural Network Toolbox Users Guide*.

BIBLIOGRAPHY

Tibshirani R, Walther G, Hastie T, 2001. Estimating the number of data clusters via the gap statistic. *Journal of the Royal Statistical Society B,* **63**: 411–423.

Tokushige S, Yadohisa H, Inada K, 2007. Crisp and fuzzy k-means clustering algorithms for multivariate functional data. *Computational Statistics* **22**: 1–16.

Trumbo B, Hudy M, Smith EP, Kim D, Wiggins BA, Nislow KH, Dolloff CA, 2010. Sensitivity and vulnerability of brook trout populations to climate change. In Carline RF, LoSapio C (eds.), *Wild Trout X: Conserving Wild Trout*, Wild Trout Symposium, West Yellowstone, Montana., 62–68.

Trumbo BA, Wise LM, Hudy M, 2012. Influence of protective shielding devices on recorded air temperature accuracy for a rugged outdoor thermal sensor used in climate change modeling. *Journal of Natural and Environmental Sciences* **3**: 42–50.

Velasco-Cruz C, Leman SC, Hudy M, Smith EP, 2012. Assessing the risk of rising temperature on brook trout: A spatial dynamic linear risk model. *Journal of Agricultural, Biological, and Environmental Statistics* **17**: 246–264.

Ver Hoef JM, Barry RD, 1998. Constructing and fitting models for cokriging and multivariate spatial prediction. *Journal of Statistical Planning and Inference* **69**: 275–294.

Wahba G, 1990. *Spline Models for Observational Data*. SIAM, Philadelphia, PA.

Wang Y, Zheng T, Zhao Y, Jiang J, Wang Y, Guo L, Wang P, 2013. Monthly water quality forecasting and uncertainty assessment via bootstrapped wavelet neural networks under missing data for harbin, china. *Environ Sci Pollut Res* **20**: 8909–8923.

Webb B, Nobilis F, 1997. A long-term perspective on the nature of the air-water relationship: a case study. *Hydrological Processes* **11**: 137–147.

BIBLIOGRAPHY

Webb BW, Clack PD, Walling DE, 2003. Water air temperature relationships in a Devon River system and the role of flow. *Hydrological Processes* **17**: 3069–3084.

Appendix A

R Code for Chapter 2

```
################Core function for fitting vclm#################
vclm <- function(x,y,t=1:length(x),K=4,lambda='GCV',...)
{
##### x, y and t must be numeric, then throw an error############
x=as.numeric(x)
y=as.numeric(y)

##### evenly distribute the knots ################
n=length(x)
k=quantile(t,probs=(1:K/(K+1)))

#######construct the splines ###################
qua=matrix(0,n,K)
for (i in 1:K)
{
```

```
for (j in 1:n)
        {
if (t[j]>k[i])
             qua[j,i]=(t[j]-k[i])^2
}
}

int=matrix(0,n,K);
for (i in 1:K)
{
int[,i]=x*qua[,i]
}

##########GCV for the optimal smoothing parameter #########
temp.lambda=0:99/10
X=cbind(rep(1,n),t,t^2,x,t*x,t^2*x,qua,int)
yfit=matrix(0,n,100)
GCV=rep(0,100)
D=diag(c(0,0,1,0,0,1,rep(1,(2*K))))

for (i in 1:100)
{
    betafit=solve(t(X)%*%X+temp.lambda[i]^2*D)%*%(t(X)%*%y)
    yfit[,i]=X%*%betafit
    L=X%*%(solve(t(X)%*%X+temp.lambda[i]^2*D)%*%t(X))
    GCV[i]=t(y-yfit[,i])%*%(y-yfit[,i])/(1-sum(diag(L))/n)^2
```

```
}

lambda=temp.lambda[which(GCV==min(GCV))]

#############Penalized Least Squares ##############
betafit=solve(t(X)%*%X+lambda^2*D)%*%(t(X)%*%y)

coef0=c(betafit[1:3],betafit[7:(6+K)])
coef1=c(betafit[4:6],betafit[(7+K):(6+2*K)])
theta0=cbind(rep(1,n),t,t^2,qua)%*%coef0
theta1=cbind(rep(1,n),t,t^2,qua)%*%coef1

list(coefficients = c(coef0,coef1),
theta0 = theta0,
theta1 = theta1,
lambda=lambda)
}

#############Define a class called vcm
vcm <- function(x, ...) UseMethod("vcm")

#############Define a default method called vcm.default###############
vcm.default <- function(x, y, t=1:length(x),K=4,lambda='GCV',...)
{
est <- vclm(x, y, t, K)
est$fitted.values <- as.vector(est$theta0 + x*est$theta1)
```

```
est$x=x
est$y=y
est$call <- match.call()
class(est) <- "vcm"
est
}

#########Define print function for the default output##############
print.vcm <- function(x, ...)
{
cat("Call:\n")
print(x$call)
cat("\nCoefficients:\n")
print(x$coefficients)
cat("\nlambda:\n")
print(x$lambda)
}

########Define the formula method##########
vcm.formula <- function(formula, data=list(), ...)
{
mf <- model.frame(formula=formula, data=data)
x <- model.matrix(attr(mf, "terms"), data=mf)
y <- model.response(mf)
est <- vcm.default(x, y, t=1:length(x),K=4,lambda='GCV',...)
est$call <- match.call()
```

```
est$formula <- formula
est
}



########Define the predict function for new data##############
predict.vcm <- function(object, newdata=NULL, ...)
{
if(is.null(newdata))
y <- fitted(object)
else{
if(!is.null(object$formula)){
x <- model.matrix(object$formula, newdata)
}
else{
x <- newdata[,1]
t <- newdata[,2]
}
y <- as.vector(object$theta0[t] + x*object$theta1[t])
}
y
}

#########Define the plot function for visualization############
plot.vcm <- function(object, ...)
{
```

```
par(mfrow=c(2,2))
plot(object$x,type='l',ylab='values',main='x and y')
lines(object$y,col=2)

plot(object$x,type='l',ylab='values',main='x and fitted y')
lines(object$fitted.values,col=2)

plot(object$theta0, type='l', ylab='intercept', main='intercept')
plot(object$theta1, type='l', ylab='slope', main='slope')
}
```

Appendix B

R Code for Chapter 3

```
library(cluster)
library(maps)
#########Generate 2 by 3 figures################
par(mfrow=c(2,3))
par(mar=c(1,1,1,1))
for (l in 1:5)
{
w=quantile(0:1)[l]

#########Calculate distance, adjusted by variogram##############
for (i in 1:62)
{
for (j in 1:62)
{
edist[i,j]=w*sum(abs((theta0[i,]-theta0[j,])/(abs(theta0[i,])+abs(theta0[j,]))))+
```

```
        (1-w)*sum(abs((theta1[i,]-theta1[j,])/(abs(theta1[i,])+abs(theta1[j,]))))
}
}

for (i in 1:62)
{
for (j in 1:62)
{
ldist[i,j]=edist[i,j]*(coef[l,1]+coef[l,2]*(1-exp(-sum(abs(sp[i,2:3]-sp[j,2:3]))
        /coef[l,3])))/(coef[l,1]+coef[l,2])
}
}

#########K-Medoids cluster by PAM##################
cl=pam(ldist,bnc[l])



#########Plot streams on a map###############
map=cbind(location,cl$cluster)
colnames(map)=c('Longitude','Latitude','group')

map.1=subset(map,map[,3]==1,select=c(Longitude, Latitude))
map.2=subset(map,map[,3]==2,select=c(Longitude, Latitude))
map.3=subset(map,map[,3]==3,select=c(Longitude, Latitude))
map.4=subset(map,map[,3]==4,select=c(Longitude, Latitude))
```

```
map('state', c('west virginia', 'virginia','north carolina','pennsylvania',
'south carolina','georgia','maryland'))

points(map.1[,1],map.1[,2],col=1,pch=19,cex=1.5)
points(map.2[,1],map.2[,2],col=2,pch=25,cex=1)
points(map.3[,1],map.3[,2],col=3,pch=22,cex=0.5)
points(map.4[,1],map.4[,2],col=4,pch=20,cex=1)

map.axes()
title('(a) Cluster map')

#########Plot intercept and slope curves###################
plot(theta0[1,],xlab='time',ylab='intercept',type='l',ylim=c(0,25),
col=cl$clustering[1],main='(b) Intercept curve')
for (i in 2:N)
{
lines(theta0[i,],col=cl$clustering[i])
}

plot(theta1[1,],xlab='time',ylab='slope',type='l',ylim=c(-0.1,.8),
col=cl$clustering[1],main='(c) Slope curve')
for (i in 2:N)
{
lines(theta1[i,],col=cl$clustering[i])
}
```

```
#########Boxplots for solar, forest and elevation#############
boxplot(solar~cl$cluster,main='(d) Solar radiation')
boxplot(forest~cl$cluster,main='(e) Percent forest')
boxplot(ele~cl$cluster,main='(f) Elevation')

#########Save#########################
wn=paste('bw',l,'.pdf',sep = "")
dev.copy(pdf,wn)
dev.off()
}
```

Appendix C

R Code for Chapter 4

```
##########Fit STVCM in the simulation study#################
a=matrix(0,35,14)

##########Generate splines for temporal effect############
t=1:365
kn=4
kt=quantile(t,prob=1:4/5)
n=365
quat=matrix(0,n,4)
for (i in 1:4)
{
for (j in 1:n)
        {
if (t[j]>kt[i])
            quat[j,i]=(t[j]-kt[i])^2
```

```
}
}

inter=cbind(rep(1,n),t,t^2,quat)

#######Generate missing data#################################
nm=120
mis=matrix(0,35,nm)

for (i in 1:35)
{
temp=sample(c(1,246),1)
mis[i,]=temp:(temp+119)
nwater[i,mis[i,]]=0
}

###########Construct design matrix and response vector############
X=matrix(0,(365*35),14)
for (i in 1:35)
{
X[((i-1)*365+1):(i*365),]=cbind(inter,inter*air[comp[i],])
}

y=rep(0,(365*35))
for (i in 1:35)
{
```

```
y[((i-1)*365+1):(i*365)]=nwater[i,]
}
X=X[which(t(nwater)!=0),]
y=y[which(t(nwater)!=0)]

#########Fit STVCM#################
h=1
lambda=0.1

for (l in 1:length(comp))
{
d=sqrt((lo[,1]-lo[l,1])^2+(lo[,2]-lo[l,2])^2)*(mis[l,1]!=mis[,1])
th=head(sort(d[d>0]),1)[1]+0.00001
d=pmax(0,(sqrt((lo[,1]-lo[l,1])^2+(lo[,2]-lo[l,2])^2)<th)*0.75*
    (1-(sqrt((lo[,1]-lo[l,1])^2+(lo[,2]-lo[l,2])^2)/(th+1))^2))/(th+1)
#d=d*(d>(th-0.00001))
#d=exp(-(u-u[l])^2/h)
Gamma=matrix(0,14,(365*35))
for (i in 1:35)
{
Gamma[,((i-1)*365+1):(i*365)]=t(cbind(inter,inter*air[comp[i],]))%*%diag(rep(d[i],365))
}
Gamma=Gamma[,which(t(nwater)!=0)]
D=diag(14)
a[l,]=solve(Gamma%*%X+lambda*D)%*%Gamma%*%y
}
```