Spatio-temporal Statistical Modeling:

Climate Impacts due to Bioenergy Crop Expansion

by

Meng Wang

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved May 2018 by the
Graduate Supervisory Committee:

Yiannis Kamarianakis, Co-Chair
Matei Georgescu, Co-Chair
Stewart Fotheringham
Mohamed Moustaoui
Mark Reiser

ARIZONA STATE UNIVERSITY

August 2018

ABSTRACT

Large-scale cultivation of perennial bioenergy crops (e.g., miscanthus and switchgrass) offers unique opportunities to mitigate climate change through avoided fossil fuel use and associated greenhouse gas reduction. Although conversion of existing agriculturally intensive lands (e.g., maize and soy) to perennial bioenergy cropping systems has been shown to reduce near-surface temperatures, unintended consequences on natural water resources via depletion of soil moisture may offset these benefits. In the effort of the cross-fertilization across the disciplines of physics-based modeling and spatio-temporal statistics, three topics are investigated in this dissertation aiming to provide a novel quantification and robust justifications of the hydroclimate impacts associated with bioenergy crop expansion. Topic 1 quantifies the hydroclimatic impacts associated with perennial bioenergy crop expansion over the contiguous United States using the Weather Research and Forecasting Model (WRF) dynamically coupled to a land surface model (LSM). A suite of continuous (2000–09) medium-range resolution (20-km grid spacing) ensemble-based simulations is conducted. Hovmöller and Taylor diagrams are utilized to evaluate simulated temperature and precipitation. In addition, Mann-Kendall modified trend tests and Sieve-bootstrap trend tests are performed to evaluate the statistical significance of trends in soil moisture differences. Finally, this research reveals potential hot spots of suitable deployment and regions to avoid. Topic 2 presents spatio-temporal Bayesian models which quantify the robustness of control simulation bias, as well as biofuel impacts, using three spatio-temporal correlation structures. A hierarchical model with spatially varying intercepts and slopes display satisfactory performance in capturing spatio-temporal associations. Simulated temperature impacts due to perennial bioenergy crop expansion are robust to physics parameterization schemes. Topic 3 further focuses on the accuracy and

efficiency of spatial-temporal statistical modeling for large datasets. An ensemble of spatio-temporal eigenvector filtering algorithms (hereafter: STEF) is proposed to account for the spatio-temporal autocorrelation structure of the data while taking into account spatial confounding. Monte Carlo experiments are conducted. This method is then used to quantify the robustness of simulated hydroclimatic impacts associated with bioenergy crops to alternative physics parameterizations. Results are evaluated against those obtained from three alternative Bayesian spatio-temporal specifications.

DEDICATION

I dedicate my dissertation to Chao, Isabella, and my beloved parents.

ACKNOWLEDGMENTS

To begin with, I would like to express my deepest appreciation to my co-advisors, Dr. Yiannis Kamarianakis and Dr. Matei Georgescu, for their scientific passion, insightful guidance, consistent support, and inspirational encouragement. They not only equip me with research skills, provide all sources of opportunities, lead me to think independently and creatively, but they also truly care about my personal life and future direction. Without their help this dissertation would be impossible.

Thanks to my committee members Dr. Mohamed Moustaoui and Dr. Stewart Fotheringham for their valuable feedback and profound suggestions on the dissertation; Dr. Mark Reiser for discussing not only research questions but questions related to graduate study. I would like to thank Melissa Wagner for our long-time collaboration, who provides great support especially in my scientific writing. I also want to acknowledge Dr. Jialun Li and Dr. Francisco Salamanca for their valuable discussion on my research. I also would like to thank all the graduate students and professors who I have worked with. Their advice and support have been instrumental to my success.

My particular appreciation would go to Dr. Alex Mahalov. It is him who brought me to step into the climate modeling and environmental statistics areas at the very beginning of my study in ASU. He is always willing to provide me opportunities, such as introducing research collaborations, supporting valuable conferences and workshops, and creating chances for interacting with important researchers in the study field.

Last but not least I would like to thank my husband Chao for the consistent companion and encouragement. I also would like to thank my parents Baomin Wang and Yuling Xiu for their tremendous support. Lastly, I would like to give my deepest love to my daughter Isabella, who makes me enjoy the beauty of life.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

Table Page

Table Page

LIST OF FIGURES

Figure Page

Figure Page

Chapter 1

INTRODUCTION

Spatio-temporal data are collected and analyzed in climatic, environmental, ecological, epidemiological and socio-economic sciences, among other research areas. Such data is usually considered to have an important statistical characteristic, namely that observations close in space and time tend to be more similar than those that are further apart (Tobler 1970; Cressie and Wikle 2015). Therefore, spatial and temporal autocorrelation should be taken into account in statistical modeling (Cressie and Wikle 2015). Ignoring spatial or temporal autocorrelation may lead to biased standard errors and artificially inflated degrees of freedom (Anselin and Griffith 1988; Wakefield 2003). Spatio-temporal model specifications vary in different circumstances, largely depending on the research questions.

With regard to time series data, detecting serially correlated monotonic trends has become a critical research question in a variety of disciplines (Vogelsang 1998; Fomby and Vogelsang 2002; Khaliq *et al.* 2009; Liebmann *et al.* 2010; Sonali and Kumar 2013; Sayemuzzaman and Jha 2014). The conventional Mann-Kendall test (Mann 1945; Kendall 1955; hereafter MK), which assumes serial independence, is one of the widely used non-parametric tests for trend-detection. However, in many real situations the data are autocorrelated: hence the conventional MK is expected to be inaccurate (Lettenmaier 1976; Khaliq *et al.* 2009; Liebmann *et al.* 2010; Sonali and Kumar 2013; Kisi and Ay 2014). Serial correlation could seriously inflate the real type I error and the power of the test, misleading conclusions related to the significance of trends, especially when sample sizes are small or moderate (Cox and Stuart 1955;

Yue and Wang 2004). To mitigate this impact, the Mann-Kendall modified trend test (Hamed and Rao 1998) and sieve bootstrap based tests (Noguchi *et al.* 2011) are considered as alternatives for serially correlated data. A number of previous studies (Kundzewicz and Robson 2004; Mudelsee 2013) have suggested the advantages of bootstrap: robustness to outliers, and avoidance of distributional assumptions are among the most important ones.

Unlike time series data which is correlated in consecutive time periods (i.e., one-dimensional), spatial dependence exists in more than two dimensions and the modeling of spatial dependence varies, depending on the type of spatial data. One type is lattice (areal) data, represented here as $\{Z(\boldsymbol{s}) : \boldsymbol{s} \in D \subset \mathbb{R}^d\}$. $Z(\boldsymbol{s})$ denotes a spatial process modeling the observations, $\boldsymbol{s}$ denotes locations, and D represents a spatial domain, considered to be discrete and fixed. Spatial associations for lattice data are typically modeled using pre-specified neighborhood structures (e.g. rook, queen, see Figure 1.1, Lloyd 2010).

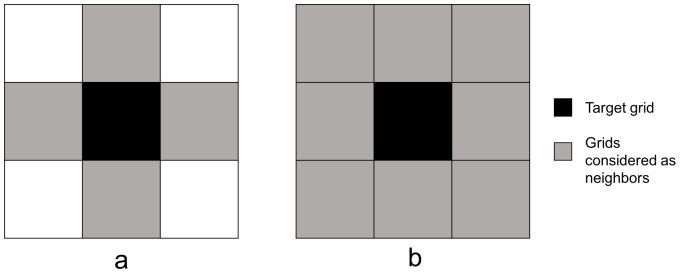

Figure 1.1: (a) Rook-type contiguity and (b) Queen-type contiguity.

In geostatistics, eigenvector spatial filtering (ESF; e.g., Griffith 2003) is a spatial modeling approach which has largely been used to capture spatial dependence of lattice data (Griffith 2010; Griffith and Paelinck 2011; Griffith 2003). ESF models spatial dependences via constructing proxy variables; these proxies are a subset of

the eigenvectors which are constructed based on the available spatial neighborhood information (Getis and Griffith 2002). To select the most appropriate candidates from a large number of eigenvectors, several strategies have been proposed and studied: stepwise regression (Griffith 2003), semiparametric modeling (Tiefelsdorf and Griffith 2007), and the least absolute shrinkage and selection operator (Lasso; Seya *et al.* 2015). Hughes and Haran (2013) extended ESF with a random effects specification (RE-ESF), which took into account spatial confounding, i.e., the proposed method mitigated the variance inflation due to the collinearity between explanatory variables and a latent spatial process (Hodges and Reich 2010; Paciorek 2010; Hughes and Haran 2013; Hanks *et al.* 2015). Murakami and Griffith (2015) further improved RE-ESF by introducing a computationally efficient REML estimation scheme and by examining the spatial scale of a spatial dependency structure.

Bayesian hierarchical models (BHM) are predominant for modeling spatial and spatio-temporal data in geostatistics and epidemiology (Cressie and Wikle 2015), when lattice data are observed in multiple time periods (hence multiple "snapshots" of a spatial process over time are available). In general, the framework of such models conditionally combines the data, the process, and the unknown parameters, to model complicated space-time processes. Specifically, the spatio-temporal covariance structure is captured using random effects. Previous studies have proposed models with different space-time structures for different purposes (i.e., Knorr-Held 2000; Li *et al.* 2012; Lee and Lawson 2016). Typically such models are estimated using advanced computational methodologies such as Markov Chain Monte Carlo (MCMC).

Despite the significant contributions of previous studies on spatio-temporal statistics, a large number of research questions remain to be answered. For instance, to the best of my knowledge, applications of eigenvector filtering are typically restricted to

the analysis of spatial data; very few applications of the technique to space-time data can be found (Patuelli *et al.* 2011; Chun 2014; Griffith and Chun 2015). In addition, the effects of misspecification when a particular BHM is chosen, are unclear. A comparative evaluation of alternative model specifications (ESF-based versus BHM) with regard to their accuracy and computational efficiency has not been presented till now. Similarly, a rather limited number of previous studies focused on how conventional trend tests compare to sieve-bootstrap-based Mann-Kendall tests.

This thesis focuses on research questions that were derived from a large spatio-temporal dataset. This dataset includes a suite of 10-year ensemble-based simulations, conducted using the Weather Research and Forecasting Model version 3.6.1 (WRF) (Skamarock et al. 2008). WRF is a nonhydrostatic model that solves the nonlinear fully compressible atmospheric equations of motion, coupled to the Noah land surface model (Noah- LSM) (Chen and Dudhia 2001; Ek *et al.* 2003). This coupling provides the capability to study the interaction of perennial bioenergy crop-induced land use change and examine hydroclimatic response to vegetation forcing (Ek *et al.* 2003). The simulations conducted for the purposes of this thesis, emphasized on the investigation of the hydroclimatic impacts due to large-scale deployment of perennial bioenergy crops across the continental United States.

1.1 Research Overview

In the effort of the cross-fertilization across the disciplines of physics-based modeling and spatio-temporal statistics, this dissertation aims to provide a novel quantification and robust justifications of the biogeophysical impacts associated with bioenergy crop expansion. State-of-the-art physics-based modeling and advanced spatio-temporal

statistical tools are used for this purpose. Specifically, the quantitative techniques that will be presented in the next sections seek to answer the following questions:

1) What are the large-scale hydroclimatic impacts associated with perennial bioenergy crop expansion over the United States?

2) Are there any statistically significant monotonic trends in regionally-averaged soil moisture?

3) Is WRF-simulated temperature impact associated with perennial bioenergy crops robust to alternative physics parameterizations?

From a methodological viewpoint, this dissertation examines 1) the robustness of WRF simulations via implementing Bayesian hierarchical models with alternative space-time structures, and 2) the effectiveness of spatio-temporal eigenvector filtering relative to BHM for environmental problems that are based on the analysis of large spatiotemporal datasets. The thesis is largely based on a trio of papers which are summarized below:

On the Long-Term Hydroclimatic Sustainability of Perennial Bioenergy Crop Expansion over the United States. The first paper, published in Journal of Climate (2017), quantifies the hydroclimatic impacts associated with perennial bioenergy crop expansion over the contiguous United States using the Weather Research and Forecasting Model dynamically coupled to a land surface model (LSM). A suite of continuous (2000–09) medium-range resolution (20-km grid spacing) ensemble-based simulations is conducted. Hövmoller and Taylor diagrams are utilized to evaluate simulated temperature and precipitation. In addition, Mann–Kendall modified trend tests and Sieve-bootstrap trend tests are used to evaluate the statistical significance of trends in soil moisture differences. Finally this research reveals potential hot spots of suitable deployment and regions to avoid.

Spatio-temporal modeling for regional climate model comparison: application on perennial bioenergy crop impacts. The second paper, published in JSM Proceedings (2016), evaluates alternative spatio-temporal Bayesian models for the analysis of WRF simulations. WRF simulated temperatures associated with control simulation bias, as well as biofuel impacts, are modeled using three spatio-temporal correlation structures. First, individual WRF simulations (see model description in Chapter 2) are modeled. Then a consensus structure, aimed at capturing spatio-temporal associations for the ensemble of WRF simulations, is discovered. The suite of WRF simulations are modeled simultaneously using the chosen consensus structure. Finally, the effects of physics parameterization on reproducing near-surface climatic conditions and the robustness of physics parameterization schemes are quantified.

Spatio-temporal modeling for regional climate model evaluation: Eigenvector filtering versus Bayesian CAR. The third paper proposes a spatio-temporal eigenvector filtering algorithm (hereafter: STEF) that takes into account spatio-temporal autocorrelation while avoiding spatial confounding. A fast estimation procedure is proposed and implemented; Monte Carlo experiments using three basic spatio-temporal structures are conducted to evaluate its performance. This proposed method is used to quantify the robustness of simulated hydroclimatic impacts associated with bioenergy crops to alternative physics parameterizations and observational datasets. Results are compared against those obtained from three alternative hierarchical Bayesian spatio-temporal specifications.

1.2 Additional published manuscripts

A realistic meteorological assessment of perennial biofuel crop deployment: A Great Plains perspective. This work, published in GCB Bioenergy, quantifies the meteorological effects of perennial bioenergy crop expansion during a normal hydrologic year (2007) and a drought year (2011) for the Southern Great Plains. This research uses realistic scenarios based on 1) field scale measurements of albedo and leaf area index (LAI) and vegetation fraction scaled according to observed albedo values, and 2) two deployment scenarios contained to marginal and abandoned lands. This study serves as a key step toward the assessment of hydroclimatic sustainability associated with perennial bioenergy crop expansion under diverse hydrometeorological conditions by highlighting the driving mechanisms and processes associated with this energy pathway (Wagner *et al.* 2017).

Sustainable Land Management for Bioenergy Crops. This work provide insights from a five-year National Science Foundation project focused on the development of spatially explicit maps of sustainable, regional "hot spots" for the large scale deployment of perennial bioenergy crops (e.g., miscanthus and switchgrass) in the United States. With environmental and economic sustainability as principal constraints, our approach integrates climate, land surface, ecosystem, and economic models. We identify "hot spots" (high suitability areas) where there is evidence of atmospheric cooling without a corresponding deterioration of water resources (e.g., significant soil moisture reduction) and simulate biomass yields on marginal lands that become inputs to our economic optimization model (Aragon *et al.* 2017).

Assessing summertime urban air conditioning consumption and its impact on anthropogenic heating in a semiarid environment. We simulate urban

air conditioning (AC) electric consumption for several extreme heat events during summertime over a semiarid metropolitan area with the Weather Research and Forecasting (WRF) model coupled to a multilayer building energy scheme. Observed total load values obtained from an electric utility company are split into two parts, one linked to meteorology (i.e., AC consumption) which is compared to WRF simulations, and another to human behavior. Built upon these results, the effect of air conditioning (AC) systems on air temperature and examines their electricity consumption for a semiarid urban environment are investigated. These studies establish a new energy consumption-modeling framework that can be applied to any urban environment where the use of AC systems is prevalent (Salamanca *et al.* 2013; 2014).

1.3 Dissertation Organization

The dissertation is structured around three manuscripts discussed in Section 1.1. Following the Introduction, Chapter 2 presents numerical simulations of perennial bioenergy crops impacts, and statistical metrics and hypothesis tests for their quantitative analysis. To further analyze the simulated data presented in Chapter 2, the following chapters focus on spatio-temporal statistical models: BHM and ESF are the main focus of Chapters 3 and 4. Directions for further research are discussed in Chapter 5.

Chapter 2

ON THE LONG-TERM HYDROCLIMATIC SUSTAINABILITY OF PERENNIAL BIOENERGY CROP EXPANSION OVER THE UNITED STATES

2.1 Introduction

Bioenergy cropping systems are increasingly recognized as a plausible and sustainable substitute for fossil fuels due to potential environmental and economic benefits (Council *et al.* 2010; Perlack *et al.* 2011). The derivation of biofuels (e.g., biobutanol, ethanol) from such cropping systems could have a number of advantages, including mitigation of climate change through greenhouse gas reduction, provision of increasing energy demands, and stabilization of energy pricing (Clifton-Brown *et al.* 2007; Campbell *et al.* 2008; Dondini *et al.* 2009; López-Bellido *et al.* 2014; Bagley *et al.* 2014; Hudiburg *et al.* 2015). Second-generation bioenergy crops (e.g., perennial grasses miscanthus and switchgrass) could serve as key alternatives to conventional feedstocks (e.g., maize) for biofuel production if planted on marginal lands (Campbell *et al.* 2008; 2013; Fargione *et al.* 2008; Field *et al.* 2008; Cai *et al.* 2010; Bagley *et al.* 2014; Hudiburg *et al.* 2016). Additionally, perennial bioenergy crops sequester carbon within the soil, and their use results in higher yields with lower nutrient input (e.g., reduced N2O) requirements relative to their annual counterparts, such as maize (Fargione *et al.* 2008; Miguez *et al.* 2008; Anderson-Teixeira *et al.* 2009; 2012; Dohleman and Long 2009; Smith *et al.* 2013; Zhuang *et al.* 2013; Gelfand *et al.* 2013; Bagley *et al.* 2014; Wagle and Kakani 2014; DeLucia 2015; Feng *et al.* 2015; Oikawa *et al.* 2015; Eichelmann *et al.* 2016; VanLoocke *et al.* 2017). Therefore, cultivating perennial

bioenergy crops could be a more sustainable approach to meet increasing energy demand and mitigate anthropogenic climate change.

While biogeochemical effects (greenhouse gas uptake and emissions) of perennial bioenergy crops have been well documented (Dondini *et al.* 2009; Gelfand *et al.* 2013; Wagle and Kakani 2014), considerable uncertainties associated with biogeophysical impacts remain (Bagley *et al.* 2014; Caiazzo *et al.* 2014; Zhu *et al.* 2017). Large-scale deployment of perennial bioenergy crops, by virtue of their transition to an altered land-use, modifies biogeophysical (e.g., direct impacts due to changes in the surface energy budget) processes. These changes could affect atmospheric boundary layer dynamics, mesoscale circulations and regional climate (Weaver and Avissar 2001; Pielke 2005; Georgescu *et al.* 2009; 2011; 2013; Mahmood *et al.* 2010; Vanloocke *et al.* 2010; Levis *et al.* 2012; Murphy *et al.* 2012). Therefore, biogeophysical impacts associated with land-use conversion to perennial bioenergy cropping systems must be considered prior to large-scale deployment.

Recent work has examined biogeophysical impacts due to landscape conversion from annual to perennial bioenergy crops, noting changes mainly attributed to higher albedo, leaf area index (LAI), and enhanced evapotranspiration (ET) (Betts 2000; Hickman *et al.* 2010; Vanloocke *et al.* 2010; Georgescu *et al.* 2009; 2011; Le *et al.* 2011; Davin *et al.* 2014; Bagley *et al.* 2014; Eichelmann *et al.* 2016; Wagle *et al.* 2016; Zhu *et al.* 2017). In addition, the importance of field-scale studies has demonstrated the significance of appropriate biogeophysical representation in process-based models that can be used to examine scenario-based environmental implications. For example, Miller *et al.* (2016), via a multi-year observational campaign, conducted field-scale measurements to determine that perennial bioenergy crops have consistently higher values of albedo than annual crops during the growing season. This higher albedo can

reduce the amount of solar energy received at the surface, affecting the partitioning of sensible, latent, and ground heat fluxes (Georgescu *et al.* 2011; 2013; Anderson-Teixeira *et al.* 2012; Anderson *et al.* 2013; Bagley *et al.* 2014; Miller *et al.* 2016). Studies have noted regional cooling (Georgescu *et al.* 2011; Le *et al.* 2011; Khanal *et al.* 2013; Goldstein *et al.* 2014; Feng *et al.* 2015) and the potential for increased precipitation (Georgescu *et al.* 2011; Khanal *et al.* 2013) associated with large-scale deployment of perennial bioenergy crops. These changes were attributable to enhanced ET due to the deeper and denser rooting systems extracting soil moisture from deeper soil depths (Vanloocke *et al.* 2010; Georgescu *et al.* 2011; Anderson *et al.* 2013; Hallgren *et al.* 2013; Ferchaud *et al.* 2015).

Changes in ET and soil moisture are directly associated with and have immediate implications for the regional hydrological cycle (Vanloocke *et al.* 2010; Georgescu *et al.* 2011; Anderson *et al.* 2013). Increased ET, owing to soil moisture depletion at deeper depths can lead to decreased surface runoff (McIsaac *et al.* 2010; Le *et al.* 2011; Wilson *et al.* 2011) and streamflow (Khanal *et al.* 2013). Concerns of surface runoff and streamflow reduction could contribute to water stress (Khanal *et al.* 2014) and have serious implications on regional water resources (McIsaac *et al.* 2010; Vanloocke *et al.* 2010; Khanal *et al.* 2013; Ferchaud *et al.* 2015).

Large-scale and long-term studies are therefore needed to better characterize hydroclimatic implications of perennial bioenergy crop expansion. For example, the previously noted cooling effect associated with perennial bioenergy crop deployment may only occur at the local and regional scale (Georgescu *et al.* 2009; 2011; Vanloocke *et al.* 2010; Hallgren *et al.* 2013). Over longer temporal scales, hydroclimatic impacts may be diminished due to natural climate variability (e.g., decadal timescale or longer). Khanal *et al.* (2014) showed that the mean increase of annual precipitation may be

smaller than the inter-annual variability of changes in precipitation when cultivating perennial bioenergy crops. Given such uncertainties, it is evident that hydroclimatic consequences of large-scale deployment of perennial bioenergy crops require further research.

Deployment of perennial bioenergy crops over abandoned and degraded lands has been proposed as a sustainable strategy (Campbell *et al.* 2008; 2013; Gelfand *et al.* 2013; Bagley *et al.* 2014; Feng *et al.* 2015). The main advantage of such an approach is avoidance of competition between food and fuel production. Few studies have assessed the implications of perennial bioenergy crops over marginal land areas, and to our knowledge, there have been no large-scale investigations to quantify hydroclimatic impacts owing to transition of abandoned and degraded farmlands to perennial bioenergy cropping systems. Here, we examine the hydroclimatic effects associated with perennial bioenergy crop deployment on abandoned and marginal land areas over the conterminous U.S. (CONUS) over a ten-year contemporary climate period utilizing a coupled land-atmosphere model. We seek to answer the following questions:

1) What are the large-scale hydroclimatic impacts associated with perennial bioenergy crop expansion?

2) Are these impacts homogeneous in space and time?

3) Can our numerical framework identify suitable hotspots of perennial bioenergy crop deployment?

By simulating deployment only over marginal or abandoned farmlands, this study portrays a more realistic depiction than previous studies for perennial-bioenergy-induced hydroclimatic consequences. This research evaluates the feasibility and long-term sustainability of large-scale deployment of perennial bioenergy crops across

CONUS while simultaneously providing a framework of feedback assessment between Land Use and Land Cover Change (LULCC) and water resources.

This chapter is arranged as follows. Section 2.2 presents a description of model configuration and experimental design, observational gridded data sets employed for model evaluation, derivation of perennial bioenergy crop expansion scenarios, and statistical methods for model evaluation and quantification of impacts. The results are presented and discussed in Section 2.3: in this section, model results are evaluated against observational data, aimed at identifying an optimal model configuration for reproducing near-surface climate conditions. Following model evaluation, hydroclimatic impacts of perennial bioenergy crop deployment are assessed. Concluding remarks and suggestions for future work are discussed in Section 2.4.

2.2 Methodology

We used the Weather Research and Forecasting model version 3.6.1 (hereafter WRF) (Skamarock 2008). WRF is a non-hydrostatic model that solves the nonlinear fully compressible atmospheric equations of motion, coupled to the Noah land surface model (Noah-LSM) (Chen and Dudhia 2001; Ek *et al.* 2003). This coupling provides the capability to study the interaction of perennial bioenergy crop-induced land use change and examine hydroclimatic response to vegetation forcing (Ek *et al.* 2003).

2.2.1 Experimental design of control simulations

Final Operational Global Analysis data (FNL) was acquired from the National Centers for Environmental Prediction for the year 2000 through the end of 2009

(NCEP 1999). FNL data are reanalysis products combining information primarily from observational weather data and Global Forecast System (GFS) model outputs, archived at a spatial resolution of one-degree by one-degree with a frequency of six hours (Research Data Archive at http://dx.doi.org/10.5065/D6M043C6). These FNL data were used to initialize and force the lateral boundaries for all WRF simulations (i.e., 2000-2009).

All simulations used a grid spacing of 20 km, consisting of 310 and 190 grid points in the east-west and north-south directions, respectively, 30 levels in the vertical direction, and a 60s time step. Numerical experiments were conducted continuously for a period of 10 years (2000 through the end of 2009), with one-month spin up (starting from Dec. 1st, 1999) to allow for land-surface conditions to reach equilibrium. Additionally, the 1-km modified IGBP MODIS 20-category land use/land cover (LULC) dataset was used to represent modern-day LULC within the Noah-LSM (Figure 2.1).

An ensemble of eight sets of control simulations (hereafter E1-E8) was conducted to determine the optimal model configuration that best reproduces near-surface climatic conditions. These ensemble members varied by choice of microphysics scheme (Hong *et al.* 2004; Lim and Hong 2010), cumulus physics scheme (Grell 1993; Grell and Dévényi 2002; Kain 2004), and utility (i.e., on or off) of spectral nudging (Miguez-Macho *et al.* 2004)(see Table 2.1). Spectral nudging corrects the systematic distortion of the large-scale flow due to the interaction with the lateral boundary conditions to derive smaller-scale processes by controlling large-scale atmospheric flow conditions in regional simulations (von Storch *et al.* 2000; Miguez-Macho *et al.* 2004). We nudged wavenumbers 0-4 in the x-direction and 0-3 in the y-direction (i.e., wavelengths longer than 1200 km) only above the boundary layer (model level equivalent to about 1500 m)

for u- and v-winds, potential temperature, and geopotential height, with a relaxation time about one hour (Table 2.1).

Table 2.1: Design of simulations. Eight Control simulations (E1-E8) that vary by choice of microphysics and cumulus physics schemes were performed. In addition, experiments with or without spectral nudging were conducted.

Ensemble member of Control simulations	Microphysics	Cumulus physics	Utilizing Spectral nudging technique
E1	WSM3	Kain–Fritsch	No
E2	WSM3	Kain–Fritsch	Yes
E3	WSM3	Grell 3D	No
E4	WSM3	Grell 3D	Yes
E5	WDM6	Kain–Fritsch	No
E6	WDM6	Kain–Fritsch	Yes
E7	WDM6	Grell 3D	No
E8	WDM6	Grell 3D	Yes

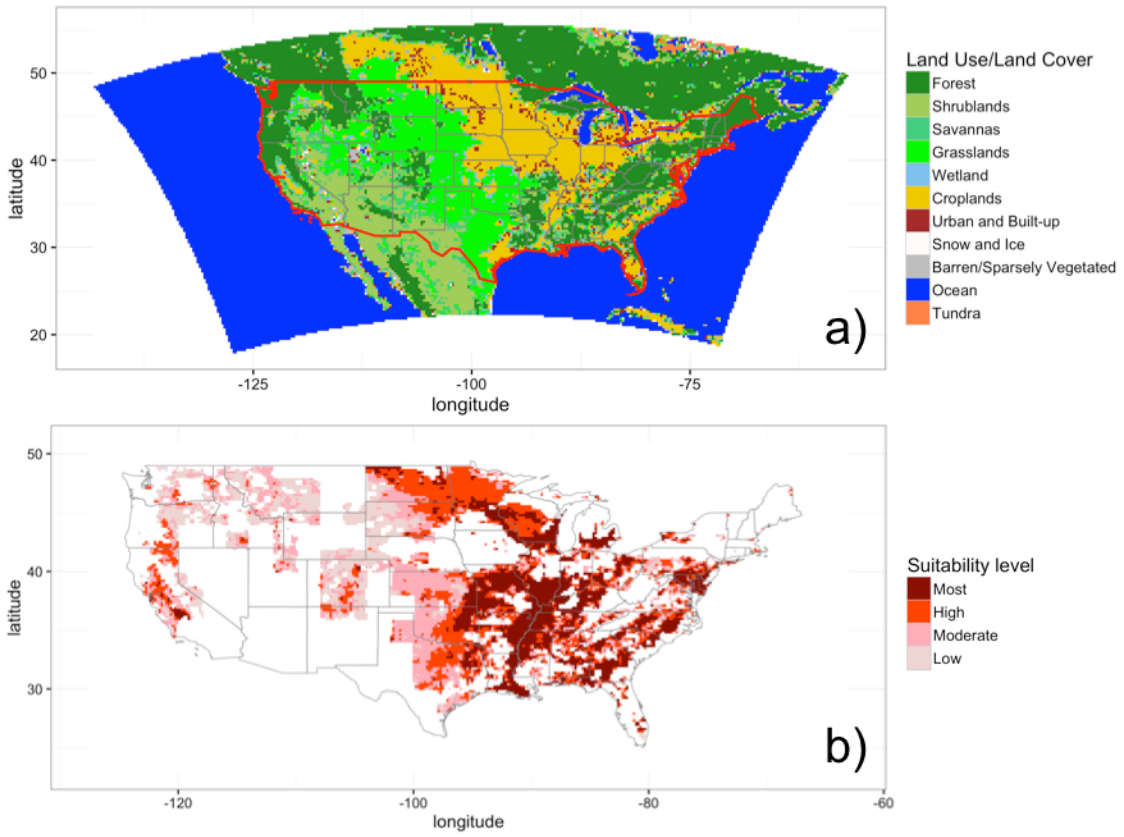

Figure 2.1: (a) Domain and MODIS landscape representation for numerical simulation experiments. Region in CONUS (outlined in red) is used for model evaluation, as well as analysis of hydroclimatic impacts associated with perennial biofuel crop deployment. (b) Suitability of perennial biofuel crops over CONUS in four quartiles. Pixels within and of suitability were reclassified as low, moderate, high and most suitable, respectively, based on Cai et al. (2011).

2.2.2 Observational Data

Two different datasets – to account for uncertainties arising from different interpolation algorithms – of gridded observational representations of temperature and precipitation were used to evaluate simulated near-surface climate. For temperature, the University of Delaware's air temperature dataset, version 3.01 (hereafter t2_DW; Willmott and Matsuura 1995) and the Global Historical Climatology Network (GHCN) and the Climate Anomaly Monitoring System (CAMS) (hereafter t2_GC; Fan and Van den Dool 2008) were utilized with a spatial resolution of 0.5°by 0.5°. Analogously, two gridded observational datasets of precipitation were used: University of Delaware Precipitation, version 3.01 (hereafter pr_DW, with the same resolution as t2_DW; Legates and Willmott 1990), and Climate Prediction Center (CPC)'s gridded Unified Gauge-Based Analysis of daily precipitation (hereafter pr_UF) with 0.25°by 0.25°longitude spatial resolution (Higgins *et al.* 2000; Chen and Knutson 2008). Datasets t2_DW, t2_GC, pr_DW, and pr_UF were provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their Web site at http://www.esrl.noaa.gov/psd/. To conduct grid cell by grid cell comparisons with simulation results, these datasets were resampled to the coarsest resolution (0.5°by 0.5°) using bilinear interpolation. Regions outside CONUS were masked out to evaluate model performance only within the study area (see Figure 2.1a).

2.2.3 Perennial bioenergy crop representation and deployment scenarios

We utilized a previously developed perennial bioenergy crop suitability dataset identifying potential areas for bioenergy crop deployment (Cai *et al.* 2010). These

data provide global suitability locations over marginal and abandoned lands using soil productivity, land slope, soil temperature, a humidity index, and additional land use information. The most realistic scenario was chosen for our study (123 million hectares available for conversion to perennial bioenergy crops throughout the U.S.), including areas of marginal mixed crop and vegetation land, grassland, savanna, and scrubland with marginal productivity, while discounting current pastureland. The original suitability data were resampled from 1km to 20km grid spacing (to match the resolution of WRF simulations) using bilinear interpolation. Suitable locations were reclassified into four suitability classes using quartile classification (i.e., low, moderate, high, and most suitable) (Figure 2.1b). Two deployment scenarios were selected using the identified suitability areas: upper 25th percentile (i.e., most suitable; hereafter Perennial25) and all suitable locations as identified by (Cai *et al.* 2010; hereafter Perennial100). Our use of both deployment scenarios was made in order to examine the largest possible range in hydroclimatic impacts associated with this bioenergy crop pathway.

Within suitable locations, perennial bioenergy crop expansion was represented via modification of relevant biophysical parameters, including albedo, LAI, and vegetation fraction (Georgescu *et al.* 2009). Albedo values were modified based on field site observation values obtained from Miller *et al.* (2016). Seasonal profiles of albedo were determined by averaging daily albedo values across two perennial plant types (switchgrass and miscanthus) and across the observed years of 2010 and 2011.

Following the phenological evolution of observed albedo, LAI and vegetation fraction values were scaled using previously reported maximum and minimum values (e.g., Dohleman and Long 2009). Albedo, LAI, and vegetation fraction values were then incorporated into Noah by taking into account latitudinal dependencies, with

shortened growing seasons to the north and lengthened growing seasons in southern regions. Specifically, albedo was depicted as:

$$\text{albedo} = \max\left\{\left[-(0.235-0.16)\left(\frac{\text{jday}-\text{centerday}}{\text{widthlai}}\right)^4 + 0.235\right], -0.16\right\} \quad (2.1)$$

where jday is the Julian day of the calendar, centerday is 197 (the assumed midpoint of the growing season and characterized as mid-July everywhere), 0.235 is the observed peak summertime albedo value and 0.16 is the observed minimum albedo value, widthlai represents the extent of the growing season in days, and is denoted as:

$$\text{widthlai} = \text{maxiwidthlai} + 0.25 \times \text{maxwidthlai} \times \frac{\text{latitude} - 30^\circ}{30^\circ - 50^\circ} \quad (2.2)$$

where maxlai = 6 (i.e., peak of the growing season), minlai = 0.1 (middle of winter when the crop is dormant), and we assume the maximum growing season LAI peaks at 30° N (i.e., maxwidthlai) and decreases linearly until 50° N, where it is equivalent to 0.75 × maxwidthlai.

Figure 2.2 shows the annual cycle of biophysical parameters for perennial bioenergy crops and existing land cover, averaged over all suitability grid cells. In general, albedo, LAI, and vegetation fraction for perennial bioenergy crops were higher than that of existing land cover from May to October. Spatial differences were apparent when examining seasonally averaged values of albedo, LAI, and vegetation fraction between Control and Perennial simulations (see Figure2.3. For albedo, the maximum difference occurs during June, July, and August (JJA). During JJA, LAI and vegetation fraction are higher over the western Plains by an average of 6 m^2 m^{-2} and 75%, respectively. Differences in biophysical characteristics were more evident for Perennial100 compared to Perennial25 simulations. It is important to mention that no bioenergy cropping systems were irrigated in this work and that no modification of default rooting depth was made.

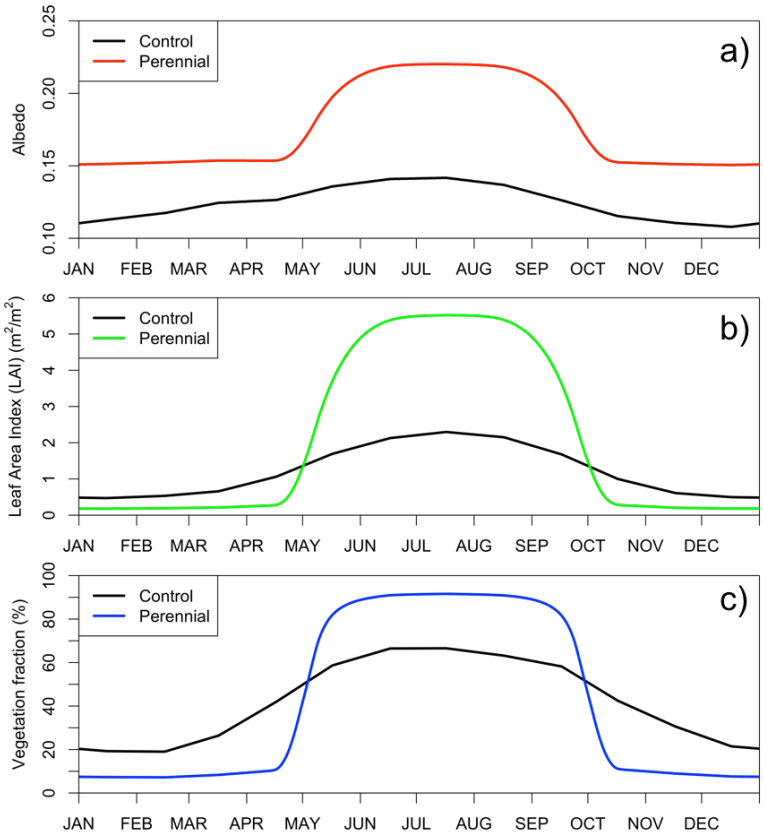

Figure 2.2: Annual cycle of biophysical representation for existing land cover and perennial bioenergy crops. Daily varying values for (a) Albedo (b) leaf area index (LAI) (m2 m-2) and (c) vegetation fraction (%) are displayed.

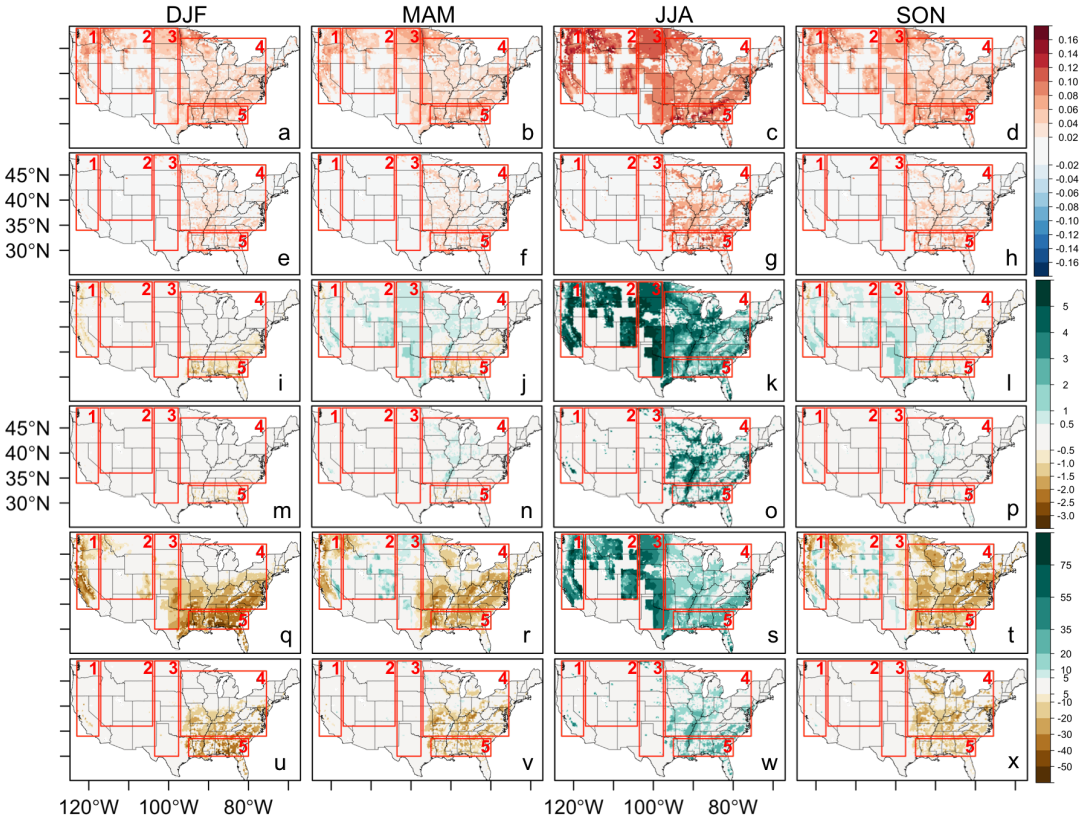

Figure 2.3: Seasonally averaged albedo difference (Perennial100-Control) for (a) DJF, (b) MAM, (c) JJA, and (d) SON. (e)-(h) Same as (a)-(d) but for Perennial25 minus Control. (i)-(p) Same as (a)-(h) but for LAI (m2 m-2). (q)-(x) Same as (a)-(h) but for vegetation fraction (%). Red rectangles outline five sub-regions for time series calculations.

Two sets of experiments were conducted over CONUS based on model skill and deployment scenarios. These experiments used the best and least skilled ensemble members (see Section 2b), based on the aforementioned model evaluation and pair of deployment scenarios (i.e., Perennial25 and Perennial100). All simulation experiments were conducted from 2000 through the end of 2009, with one month of spin-up in Dec. 1999, to allow the land surface state to equilibrate (see Table 2.2).

Table 2.2: List of bioenergy crop sensitivity simulations.

WRF Simulation	Scenario	Spin Up	Analysis Time
Control	Control_E1 Control_E8	Dec 1-31, 1999	Jan 1, 2000- Dec 31, 2009
Perennial bioenergy crop deployment	Perennial100_E1 Perennial25_E1 Perennial100_E8 Perennial25_E8		

2.2.4 Exploratory Statistics for WRF-Model Evaluation

Hovmöller and Taylor diagrams were utilized to evaluate simulated temperature and precipitation. Hovmöller diagrams (Hovmöller 1949) visually summarize model performance over space and time in two dimensions. More specifically, information is averaged across latitude bands and displayed on the X-axis (i.e., longitudinal dimension is removed) whereas time is represented on the Y-axis. These diagrams were used to quantify monthly averaged relative differences (i.e., dimensionless values) between the eight control simulations (E1 through E8) and the aforementioned gridded observation datasets (differences were normalized by the corresponding observations).

The steps for creating a Hovmöller diagram are as follows. The simulated bias of 2m temperature Control_E1 relative to observation DW from 2000 to 2009 is used as an example:

1) The boundaries of the area to analyze was specified. In this case, it was the continental U.S.

2) Time intervals were specified for data calculation: since we were interested in capturing monthly cycles, the simulated bias was aggregated (i.e. monthly averaging) from three-hour frequency to monthly frequency.

3) Starting from January 2000, we averaged the magnitudes of simulated bias in all the grid boxes (pixels) across each latitude strip.

4) Repeat Step 3 for every month from 2000 to 2009. Each monthly simulated bias corresponds to one row in the Hovmöller diagram.

Additionally, Taylor diagrams were used to summarize simulation skill based on seasonally averaged differences between each ensemble member (i.e., test field (t)) and observed 2m temperature or precipitation (i.e., reference field (r)). More generally speaking, this diagram can statistically quantify the degree of similarity between two fields. Taylor diagrams illustrate the variances of the test and reference fields (ρ_t^2 and ρ_r^2, respectively), the centered RMS difference between the fields (E′), and correlation coefficients between the test and reference fields (ρ), simultaneously in one diagram (Taylor 2001). Mathematically, the three statistics are related by the following formula:

$$
\begin{aligned}
E'^2 &= \sigma_r^2 + \sigma_t^2 - 2\sigma_r\sigma_t\rho \\
\text{where: } \rho &= \frac{\frac{1}{N}\sum_{n=1}^{N}\left(t_n - \overline{t}\right)\left(r_n - \overline{r}\right)}{\sigma_t\sigma_r} \\
E' &= \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left[\left(t_n - \overline{t}\right)\left(r_n - \overline{r}\right)\right]^2} \\
\sigma_t^2 &= \frac{1}{N}\sum_{n=1}^{N}\left[\left(t_n - \overline{t}\right)\right]^2 \\
\text{and } \sigma_r^2 &= \frac{1}{N}\sum_{n=1}^{N}\left[\left(r_n - \overline{r}\right)\right]^2
\end{aligned}
\tag{2.3}
$$

Based on Hovmöller and Taylor diagram metrics, two of the eight control ensemble members were selected as the most and least skillful, respectively, and served as baseline simulations using the existing land cover (hereafter Control), against which simulations representing perennial bioenergy crop expansion were compared. Incorporation of

bioenergy crops (see Section 2.2.3) was made for both sets of model parameterization options (i.e., corresponding to the most and least skillful ensemble members) to examine whether the sensitivity to landscape change, and if so to what extent, depends on simulation skill.

2.2.5 Trend Tests for Serially Correlated Data

To assess the sustainability of perennial bioenergy crop expansion, the Mann-Kendall modified trend test (for seasonal time series in the presence of serial correlation, Hamed and Rao (1998) was used. Sieve bootstrap based tests (Noguchi *et al.* 2011) was also used to evaluate statistical significance of trends in soil moisture differences, when serial autocorrelation is included. These tests provide evidence on the possibility of existence of a monotonic upward or downward (not necessarily linear) trend of soil moisture depletion over time. Conventional Mann-Kendall tests form the basis of the two aforementioned trend tests; the null hypothesis states that that there is no significant trend (i.e., independent data), whereas the alternative hypothesis supports the existence of a (not necessarily linear) trend.

The Mann-Kendall Statistic S contains the information of net increments or decrements of a time series:

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \operatorname{sgn}\left(x_j - x_i\right) \tag{2.4}$$

where n represents the number of measurements; x_i and x_j denote the i^{th} and the j^{th}

observation, respectively, and sgn() is the sign function defined as:

$$\text{sgn}(x_j - x_i) = \begin{cases} 1, & \text{if } x_j - x_i > 0. \\ 0, & \text{if } x_j - x_i = 0. \\ -1, & \text{otherwise.} \end{cases} \tag{2.5}$$

S is approximately normally distributed when $n > 10$; the mean of S is zero and its variance can be calculated by:

$$\text{var}(S) = \frac{1}{18}\left[n(n-1)(2n+5) - \sum_{i=1}^{m} t_i(t_i-1)(2t_i+5)\right] \tag{2.6}$$

A set of data that has the same value is a tied group. In Eq.(2.6), m is the number of tied groups, each with tied observations. The test statistic Z can be calculated by:

$$Z = \begin{cases} \dfrac{S-1}{\sqrt{var(S)}}, & \text{if } S > 0, \\ 0, & \text{if } S = 0, \\ \dfrac{S+1}{\sqrt{var(S)}}, & \text{otherwise.} \end{cases} \tag{2.7}$$

Z follows the standard normal distribution (Hamed and Rao 1998) asymptotically. A positive value of Z indicates an upward trend, and a negative value indicates a downward trend.

To take into account the autocorrelation structure of the time series data, Hamed and Rao (1998) investigated a modified Mann-Kendall test using adjusted variance based on effective sample size:

$$\text{var}(S)^{\star} = \text{var}(S)\frac{n}{n_s},$$
$$\text{where } \frac{n}{n_s} = 1 + \frac{2}{n(n-1)(n-2)}\sum_{i=1}^{p}(n-i)(n-i-1)(n-i-2)p_s(i) \tag{2.8}$$

In Eq.(2.8), n denotes the number of observations in the time series, while n_s represents the effective sample size which essentially accounts for serial correlation in the data. It can be seen that n/n_s will be greater than 1, less than 1, and equal to 1 when data are positively, negatively, and not autocorrelated (no adjustment is made), respectively. $p_s(i)$ is the lag-i autocorrelation between ranks of the observations, where p is the maximum time lag under consideration.

Alternatively, Sieve-bootstrap for the Mann-Kendall tests (Noguchi *et al.* 2011) estimate the trend after the time series data has been prewhitened. More specifically, the autoregressive structure AR(p) of a time series $(y_1, \ldots, y_T)$ is estimated under the null hypothesis of no trend in the first stage:

$$y_t = \sum_{k=1}^{p} \hat{\alpha}_k y_{t-k} + \hat{\epsilon}_t \tag{2.9}$$

In (2.9), $\hat{\alpha}_k, k = 1, \ldots, p$ represent sample estimates of population autoregressive parameters. By removing autocorrelation component from the original data, the obtained residuals $(e_1, \ldots, e_T)$ were bootstrapted for constructing resampled residuals as a generating noises. By adding together the generating noises and simulated new time series based on AR(p), B sieve bootstrap samples, denoted by $y^*_{it}, i = 1, \ldots, B$ were constructed. Using this method, the new sieved bootstrapped time series contain similar serial dependence structure to the original data. Then the Mann-Kendall modified trend test statistic can be calculated for each sieve-bootstrapped time series.

In this study, spatially averaged time series of soil moisture differences were aggregated from daily to monthly frequency to conduct the trend tests. To compensate for the number of inferences, a Bonferroni adjustment was applied using a higher significance threshold for individual comparisons. Specifically, test-specific p-values smaller than 0.001 characterized statistical significance in order to achieve a family-wise Type I error rate (false positives) approximately equal to 0.05.

2.3 Results

2.3.1 Model Evaluation

In general, model skill was superior for temperature compared to precipitation across all simulated years and ensemble members. Hovmöller diagrams (Figs. 2.4, 2.5) show minimal variability in simulated near-surface temperature (i.e., at 2 meters above ground), but high variability for precipitation across ensemble members. Monthly averaged temperature biases were small compared to both observational datasets (Fig. 2.4). However, temperatures biases varied according to latitude and time of year. During summer, simulated temperatures exhibited a positive bias primarily over southern areas, whereas during winter, simulated temperatures exhibited a negative bias over northern locations. Ensemble members E4 and E8 performed best in simulating temperature especially during summer, whereas ensemble members E1 and E5 exhibited the largest warm bias (Figure 2.4). Overall, ensemble member E8 (see Table 2.2; with Microphysics WDM6) produced the best correspondence to wintertime temperatures while demonstrating minimal summertime warm biases, whereas ensemble member E1 (with microphysics WSM3) displayed the largest underestimate of near-surface temperatures.

Unlike temperature, monthly averaged simulated precipitation biases were highly variable. Fig. 2.5 shows normalized precipitation differences generally up to 5 times greater than observed precipitation, which was more prevalent when compared with the second observed dataset. Additionally, precipitation biases were greater over latitudinal belts below 30°N or above 45°N. The disparity between simulated precipitation and observation datasets is largely explained by the different algorithms utilized to create

the gridded observational datasets themselves. Despite this disparity, ensemble members E4 and E8, which used the Grell-3D cumulus scheme and spectral nudging, performed better than the other ensemble members (Fig. 2.5). Ensemble members E1 and E5, which used the Kain-Fritsch cumulus scheme without spectral nudging, performed worse.

In addition to evaluating Hovmöller diagrams, Taylor diagrams (which permit simultaneous assessment of multiple statistical metrics) also show high model skill in simulating temperature, but only moderate model skill for precipitation (Fig. 6). For near-surface temperature, considerable clustering among all ensemble members is evident, indicative of reduced near-surface temperature sensitivity to choice of model physics (Fig. 2.6a-d). All ensemble members show similar standard deviation, correlation coefficients near 0.96, and centered RMSE ranging from 0.25 to 0.4 °C.

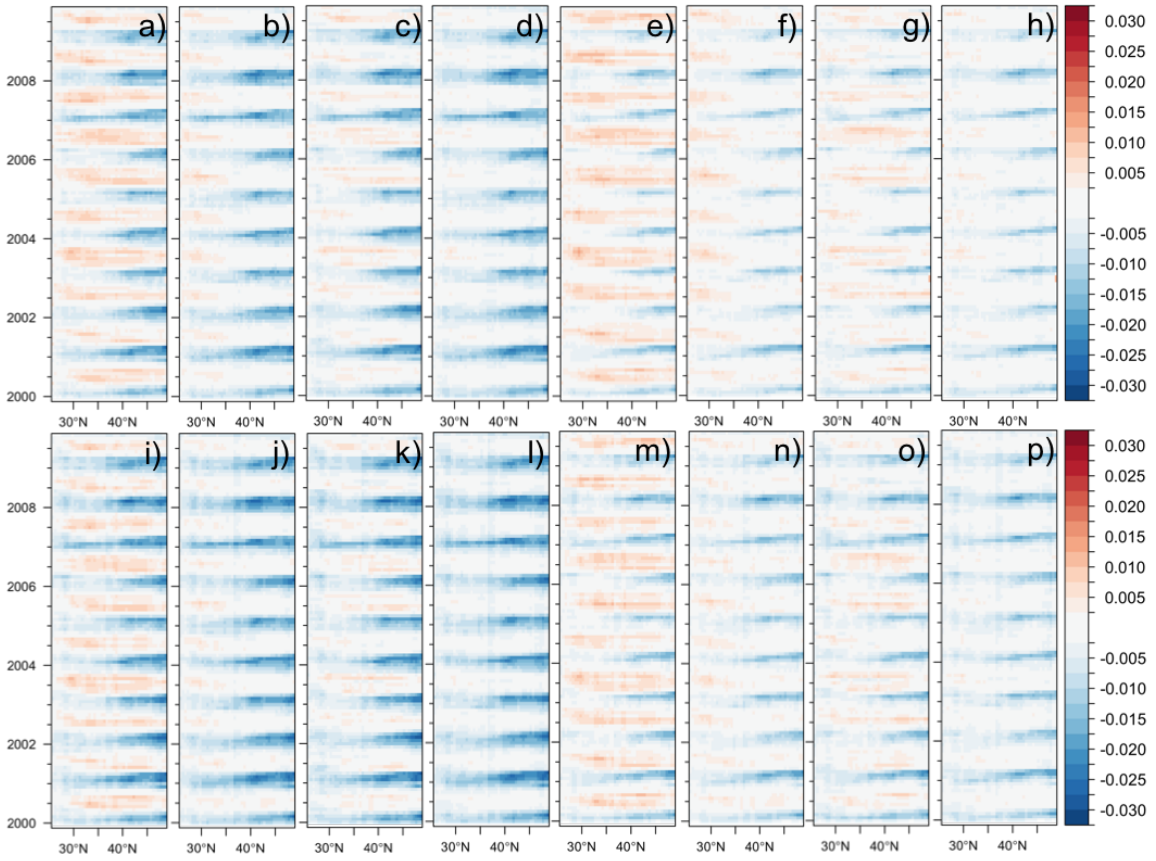

Figure 2.4: Hovmöller diagrams of monthly averaged relative differences of near-surface temperature (K) (relative differences were derived by subtracting observations from control simulations, and then dividing by the corresponding observations) between (a)-(h) the eight control simulations (E1-E8) and the observational dataset t2_DW. (i)-(p) Same as (a)-(h) but for the observational dataset t2_GC.

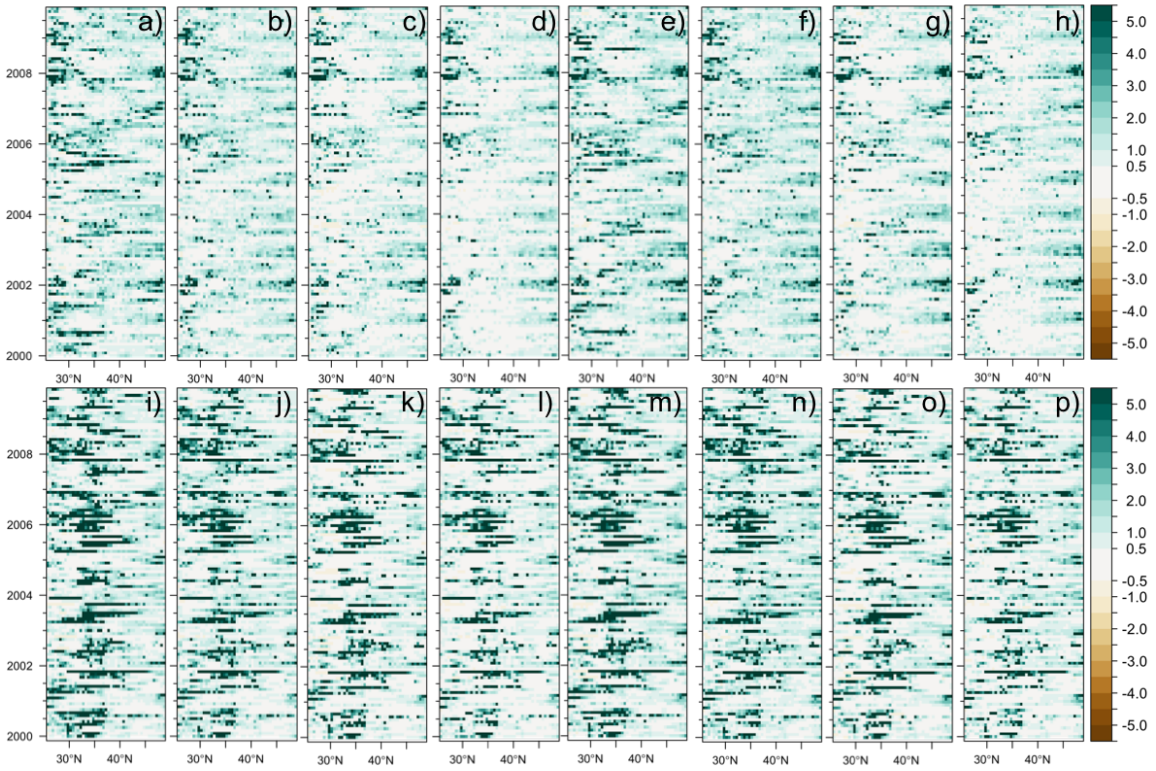

Figure 2.5: Hovmöller diagrams of monthly averaged relative differences of precipitation (mm d-1) (relative differences were derived by subtracting observations from control simulations, and then dividing by the corresponding observations) between (a)-(h) the eight control simulations (E1-E8) and the observational dataset pr_DW. (i)-(p) Same as (a)-(h) but for the observational dataset pr_UF.

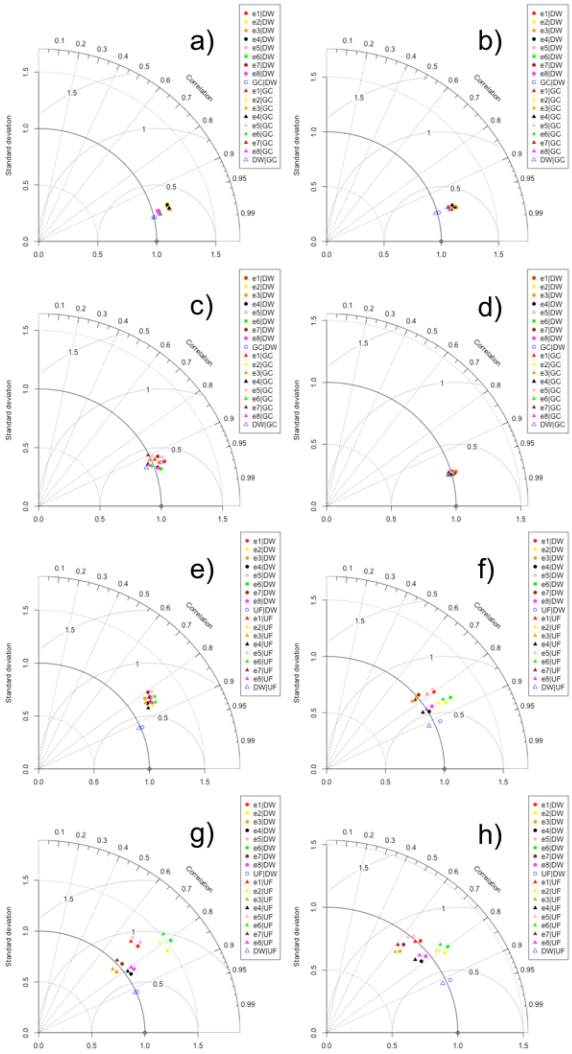

Figure 2.6: Taylor diagrams of seasonally averaged near surface temperature between observations and control simulations over 10 years (2000-2009) in (a) DJF, (b) MAM, (c) JJA, and (d) SON. (e)-(h) Same as (a)-(d) but for precipitation. Dots represent simulation skill relative to observed dataset of University of Delaware Air Temperature and Precipitation (i.e., DW), whereas triangles correspond to observed temperature and precipitation datasets of GHCN_CAMS Gridded 2m Temperature and CPC U.S. Unified Precipitation (i.e., GC and UF), respectively. Hollow symbols represent the relationship between gridded observational datasets. Correlation coefficients between modeled and observed variables are shown in angular axes. Normalized standard deviation and centered root mean square error (RMSE) are proportional to the distance from the origin and the (1,0) point, respectively.

For simulated precipitation, considerable spread among the ensemble members is evident, indicating enhanced sensitivity to the choice of physics parameterizations employed here (Fig.2.6e-h). The standard deviation of the simulated precipitation values was 0.9 to 1.5 times greater than that of the observations. Centered RMSE values ranged between 0.5 to 1 mm d-1. Correlation coefficients for all ensemble members were lowest during summer and fall (generally between 0.65 to 0.8), coinciding with the period of time when large-scale synoptic forcing is absent and precipitation is convectively driven. Nevertheless, ensemble members E4 and E8 consistently performed better than other members, especially during the convective season, exhibiting correlation coefficients in excess of 0.8, lowest standard deviation ratio of 1 relative to that of observations, as well as lowest centered RMSE of 0.7. Ensemble members E1 and E5 had the least model skill in simulating precipitation; this was especially evident during the summer (e.g., these ensemble members had a lowest correlation coefficient of 0.65).

Based on the aforementioned results, ensemble member E8, which used the Grell-3D cumulus scheme, WDM6 microphysics parameterization and spectral nudging turned on, performed the best, whereas, ensemble member E1, which used the Kain-Fritsch cumulus scheme, WSM3 microphysics parameterization and spectral nudging turned off, performed the worst. In the following analysis, ensemble members E8 and E1 were identified as the best and least skilled members, respectively. Both ensemble members (i.e., E8 and E1) were modified to incorporate bioenergy crops (see Section 2c) to assess whether the sensitivity to landscape change depends on simulation skill, and if so to what extent.

2.3.2 Hydroclimatic impacts

2.3.2.1 Temperature

We present results as differences in 10-year seasonally-averaged hydroclimatic variables between the perennial bioenergy crop simulations and the contemporary landscape utilized in control simulations. Overall, seasonal averages of near-surface temperature differences illustrate cooling associated with deployment of perennial bioenergy crops (Fig.2.7). Maximum simulated cooling occurs during the peak of perennial bioenergy crop greenness (JJA) for all deployment scenarios. During this period, near-surface temperature decreases dramatically over the southern Great Plains with maximum cooling on the order of 5 °C for the full deployment scenario (i.e. Perennial100_E1 and Perennial100_E8, corresponding to Fig.2.7c and 2.7g). The Pacific Coast, western mountains (designated as regions 1 and 2, respectively) exhibit moderate temperature decreases of approximately 2-4 °C. This cooling is gradually attenuated from the Central Plains to the Northeast U.S. (i.e., within regions 4 and 5, respectively). Under the reduced deployment scenario (i.e., Perennial25_E1 and Perennial25_E8), near surface cooling associated with perennial bioenergy crop deployment is more localized and primarily restricted to regions 4 and 5. Within these regions, the maximum cooling is restricted to approximately 3 °C during summer months (Fig. 2.7i-p). Only minimal differences in simulated cooling were evident when comparing ensemble member E1 and E8 results (i.e., compare Fig.2.7c and 2.7g), indicating that the simulated near-surface temperature sensitivity to bioenergy crop deployment was independent of model performance.

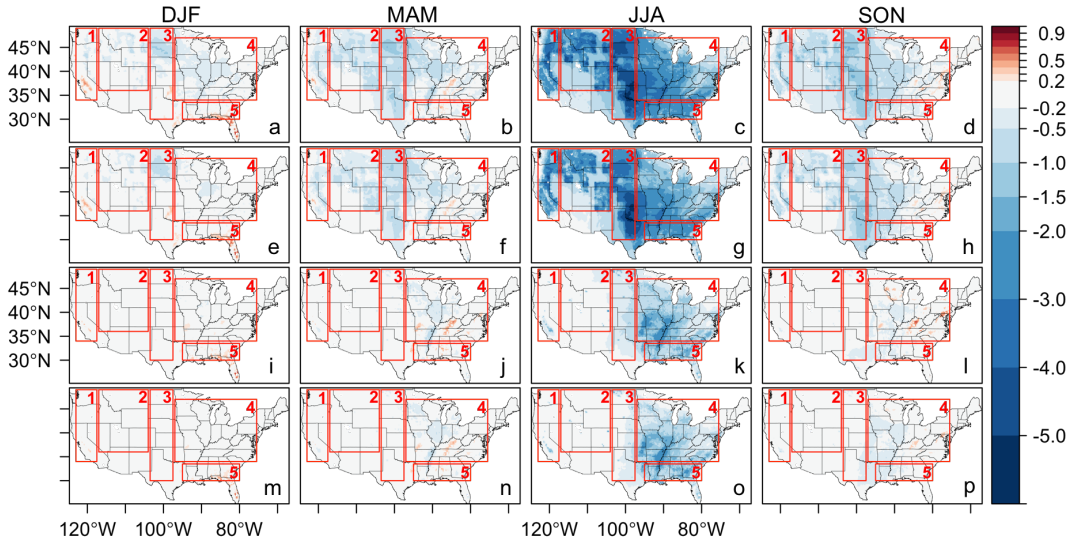

Figure 2.7: Seasonally averaged near-surface temperature difference (°C) (Perennial100_E1-Control_E1) over one decade (2000-2009) for (a) DJF, (b) MAM, (c) JJA, and (d) SON. (e)-(h) Same as (a)-(d) but for difference of Perennial100_E8 minus Control_E8. (i)-(l) Same as (a)-(d) but for difference of Perennial25_E1 minus Control_E1. (m)-(p) Same as (a)-(d) but for difference of Perennial25_E8 minus Control_E8. Red rectangles outline five sub-regions for time series calculations.

To better examine hydroclimatic impacts over time, time-series plots of temperature differences are calculated for each of the five sub-regions depicted in Fig. 2.8. These sub-regions include the Pacific Coast (sub-region 1), western mountains (sub-region 2), western Great Plains (sub-region 3), central/eastern U.S. (sub-region 4), and Gulf Coast (sub-region 5). Across all sub-regions, cooling occurs from May to October, coinciding with the higher albedo of perennial bioenergy crops (Fig.**??**a). Under the full deployment scenario, maximum cooling ranges between 3-5 °C over region 3 (i.e., western Great Plains), whereas, regions 4 (central/eastern U.S.) and 5 (Gulf Coast) illustrate a maximum cooling ranging between 1-2 °C under the reduced deployment scenario. In terms of ensemble member performance, E8 and E1 overlap considerably. However, E8 displays less variability in annual cycle differences as indicated by the narrower standard deviation band when compared to E1 (Fig. 2.8). Despite this small difference, uncertainty due to model physics parameterization is secondary to the simulated signal of cooling impact. Moreover, we consider the simulated thermal impacts robust as temperature differences and the associated annual variability consistently exhibits cooling, with only small exceptions evident for reduced deployment experiments for some regions (e.g., region 1).

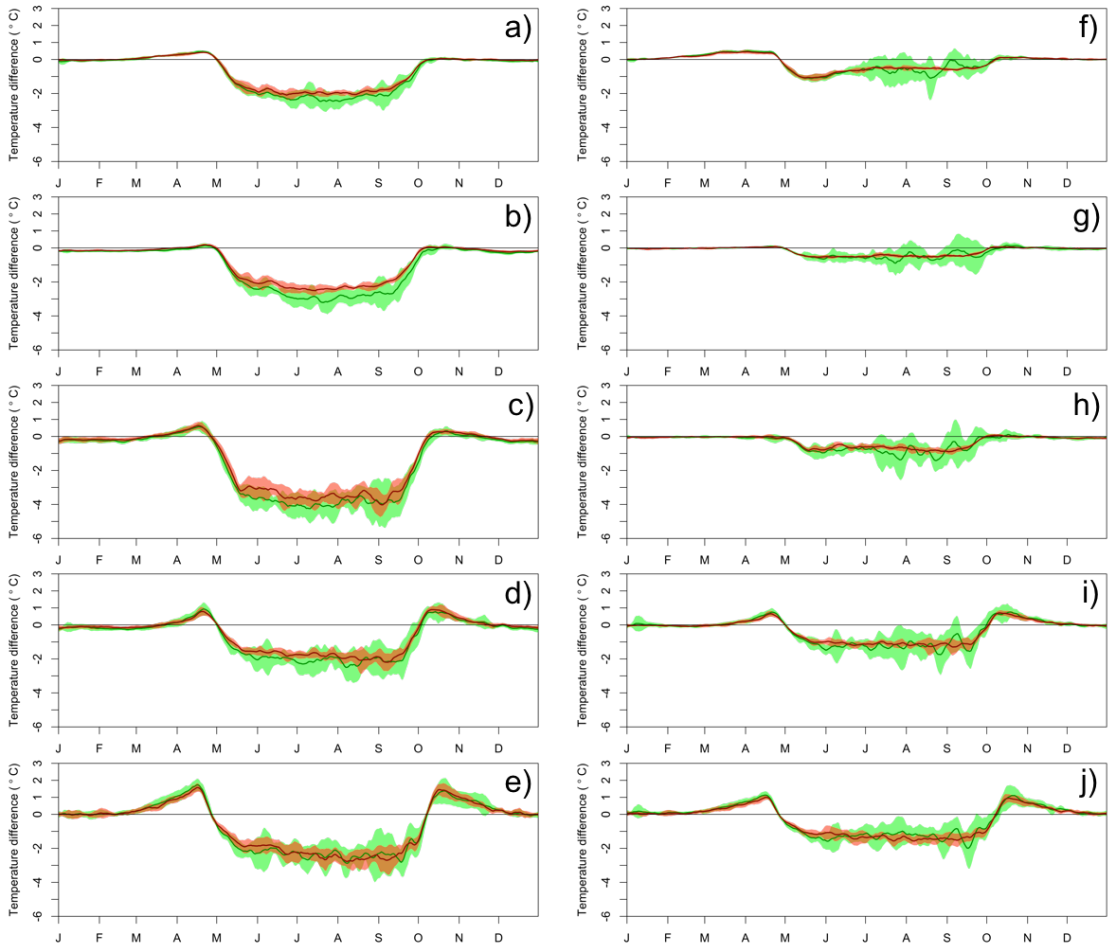

Figure 2.8: Annual cycle of surface temperature differences (°C), averaged only over grid cells undergoing land surface modification under Perennial100 scenario (a) region 1, (b) region 2, (c) region 3, (d) region 4, and (e) region 5. (f)-(j) Same as (a)-(e) but under Perennial25 scenario. Green and red lines indicate averaged annual cycle of simulated impact over the decadal period using ensemble member E1 and E8, respectively. Bands of one standard deviation above and below the mean annual cycle are shaded with the corresponding color.

2.3.2.2 Surface energy balance

Similar to simulated temperature patterns, sensible heat flux associated with perennial bioenergy crops also decreases under both deployment scenarios (see Fig.2.9). This decrease is maximized during the summer months especially under the full deployment scenario. Under this scenario, peak reduction in sensible heat flux, ranging between 40-70 W m-2 was evident over western and central portions of the U.S. (regions 1, 2, and 3). Under the reduced deployment scenario, the reduction in sensible heat was moderated to only 20 W m-2. This reduction was most noticeable in the central/eastern U.S. and Gulf Coast areas (regions 4 and 5), unlike the full deployment scenario, which exhibited greatest decrease in sensible heat along or west of the 100^{th} meridian.

The temporally varying nature of sensible heat flux differences for the individual sub-regions also indicates lower sensible heat fluxes associated with perennial bioenergy crops during the growing season (Fig.2.10). For regions 1-3, the greatest decrease occurs from May to mid-June. In regions 4-5, sensible heat flux is more gradually reduced and remains nearly constant for the majority of the growing season. The reduction in sensible heat flux for regions 4 and 5 coincide with reduced temperature differences for these two regions. Under the full deployment scenario, sensible heat decreases by a maximum of 45 W m-2 in region 3. Under the reduced deployment scenario, the decrease in sensible heat is minimized to 15-25 W m-2.

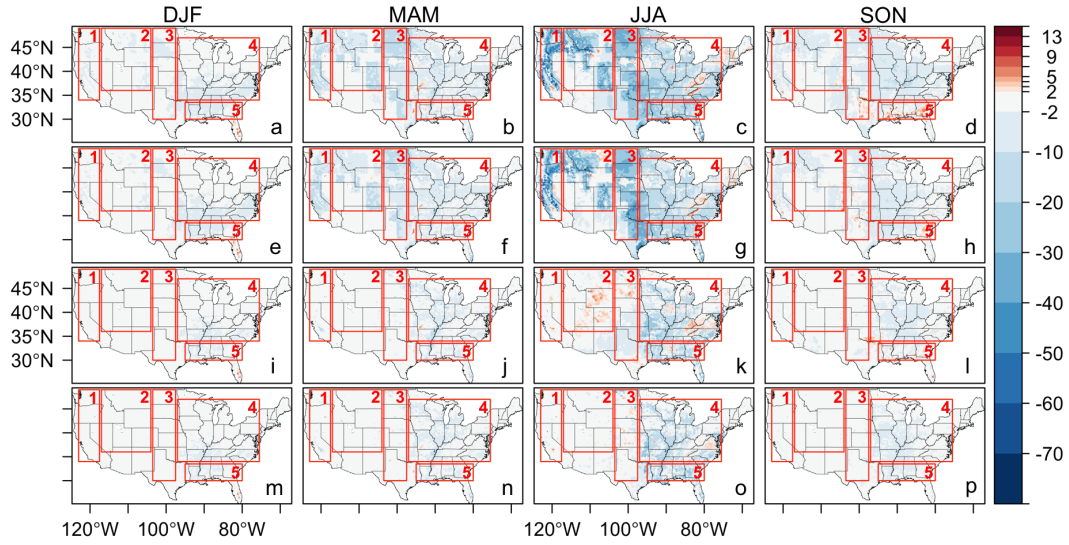

Figure 2.9: Seasonally averaged sensible flux difference (W m-2) (Perennial100_E1 - Control_E1) over one decade (2000-2009) for (a) DJF, (b) MAM, (c) JJA, and (d) SON. (e)-(h) Same as (a)-(d) but for difference of Perennial100_E8 minus Control_E8. (i)-(l) Same as (a)-(d) but for difference of Perennial25_E1 minus Control_E1. (m)-(p) Same as (a)-(d) but for difference of Perennial25_E8 minus Control_E8. Red rectangles outline five sub-regions for time series calculations.

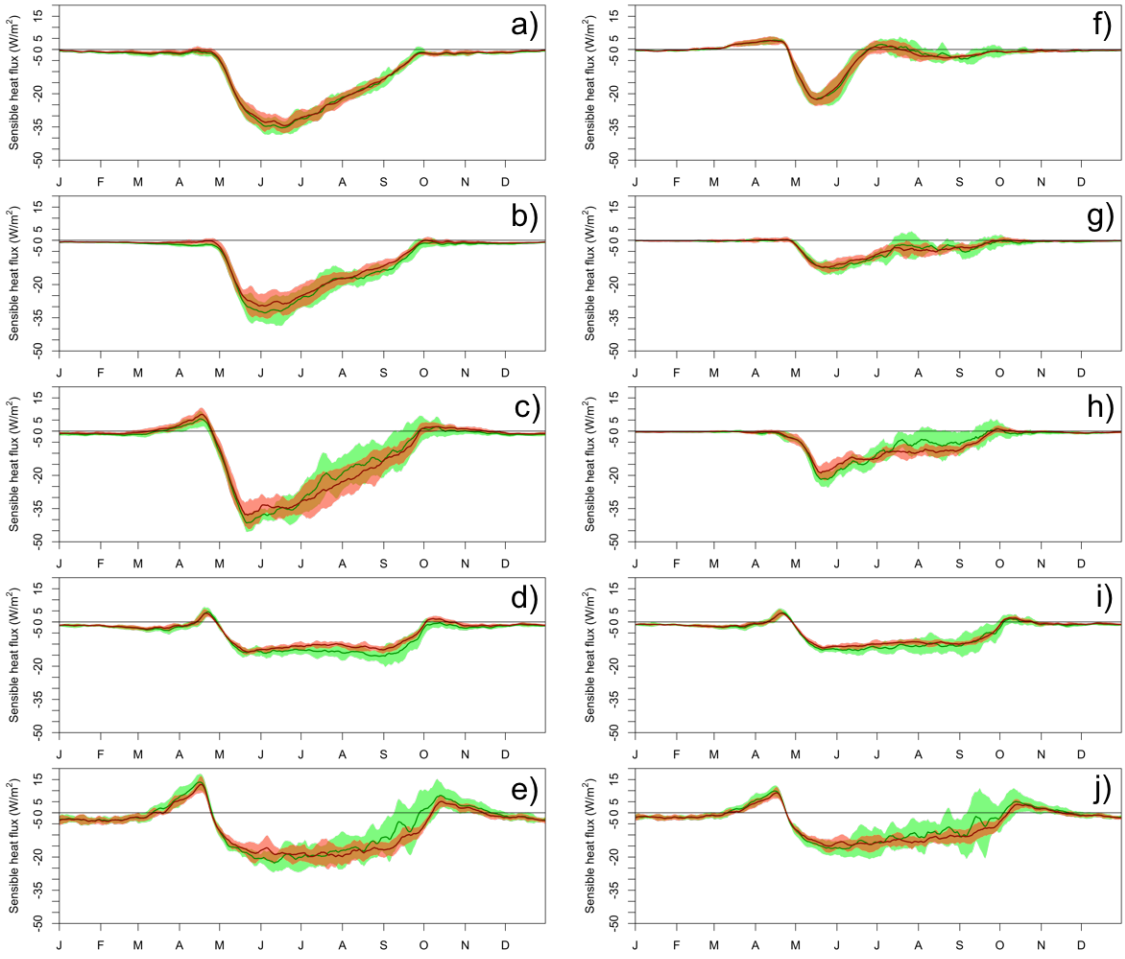

Figure 2.10: Annual cycle of sensible heat flux difference (W m-2) averaged only over grid cells undergoing land surface modification under Perennial100 scenario (a) region 1, (b) region 2, (c) region 3, (d) region 4, and (e) region 5. (f)-(j) Same as (a)-(e) but under Perennial25 scenario. Green and red lines indicate averaged annual cycle of simulated impact over the decadal period using ensemble member E1 and E8, respectively. Bands of one standard deviation above and below the mean annual cycle are shaded with the corresponding color.

Despite consistent decreases in sensible heat, latent heat fluxes associated with perennial bioenergy crop expansion exhibits geographically dependent changes (Fig. 2.11). During the growing season, latent heat fluxes increase, by up to 55 W m-2, over Pacific Coast, western mountains, and western Great Plains regions (regions 1, 2, and 3) under the full deployment scenario. However, over eastern portions of the U.S. (regions 4 and 5), latent heat fluxes decrease, generally between 15-25 W m-2 for full and reduced deployment scenarios. In addition, according to time series plots of latent heat flux differences (Fig. 2.12), regions 1, 2, and 3 display higher latent heat fluxes associated with perennial bioenergy crops through early portions of the summer, followed by a gradual decrease until October. Over regions 4 and 5, latent heat flux differences are small during the growing season. Notably, decreases in latent heat fluxes are evident from April to May, and October to November, coinciding with lower LAI and vegetation fraction values for perennial bioenergy crops relative to the existing land cover (see Fig. 2.2b-c).

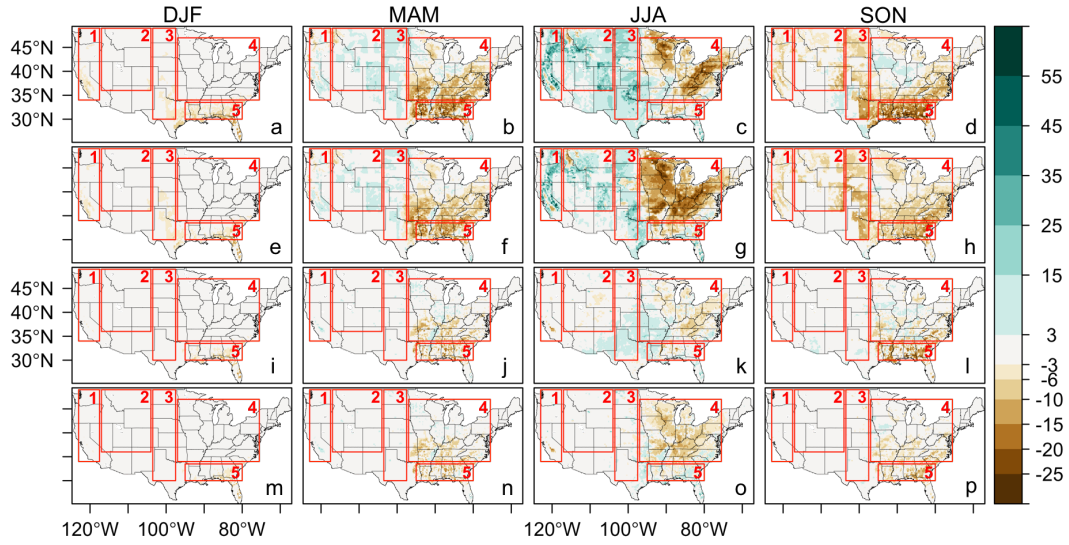

Figure 2.11: Seasonally averaged latent heat flux difference (W m-2) (Perennial100_E1 - Control_E1) over one decade (2000-2009) for (a) DJF, (b) MAM, (c) JJA, and (d) SON. (e)-(h) Same as (a)-(d) but for difference of Perennial100_E8 minus Control_E8. (i)-(l) Same as (a)-(d) but for difference of Perennial25_E1 minus Control_E1. (m)-(p) Same as (a)-(d) but for difference of Perennial25_E8 minus Control_E8. Red rectangles outline five sub-regions for time series calculations.

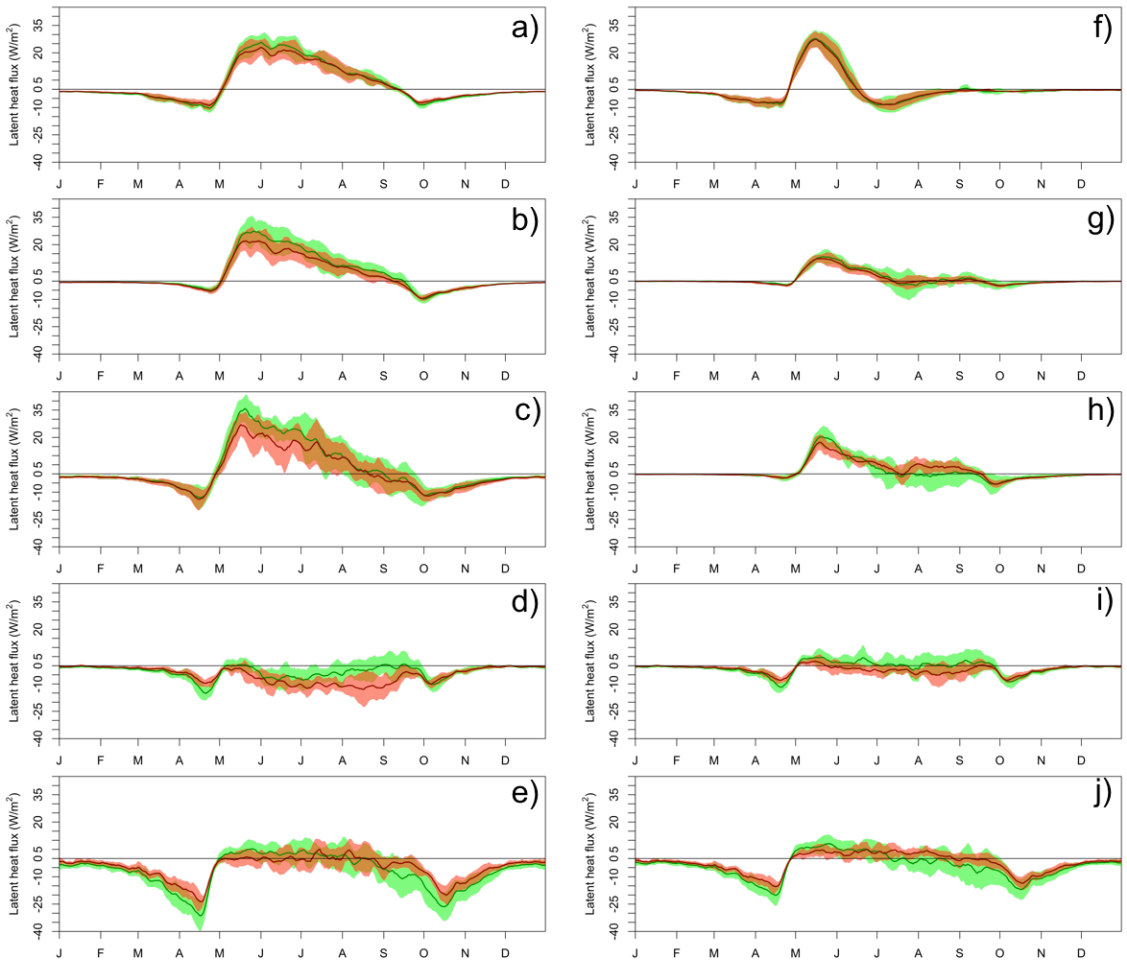

Figure 2.12: Annual cycle of latent heat flux difference (W m-2) averaged only over grid cells undergoing land surface modification under Perennial100 scenario (a) region 1, (b) region 2, (c) region 3, (d) region 4, and (e) region 5. (f)-(j) Same as (a)-(e) but under Perennial25 scenario. Green and red lines indicate averaged annual cycle of simulated impact over decadal period using ensemble member E1 and E8, respectively. Bands of one standard deviation above and below the mean annual cycle are shaded with the corresponding color.

2.3.2.3 Soil moisture

Changes in soil moisture associated with perennial bioenergy crops are inversely related with latent heat flux changes. Soil moisture changes are evident in both shallow (10 - 40 cm; Fig. S1) and deeper (40 - 100 cm; Fig. 2.13) soil depth levels. Under the full deployment scenario, soil moisture was reduced over western and central portions of the U.S. (regions 1, 2, and 3) during summer and fall. Within these regions, volumetric soil moisture decreased by up to 0.17 m3 m-3 and 0.20 m3 m-3 for shallow and deeper soil depths, respectively. In the central/eastern U.S. (region 4), unlike other regions, soil moisture increased by up to 0.07 m3 m-3 and 0.10 m3 m-3 for shallow and deeper soil depths, respectively. Soil moisture differences were minimal under the reduced deployment scenario with minor changes manifested in regions 4 and 5, respectively ($<$0.05 m3 m-3).

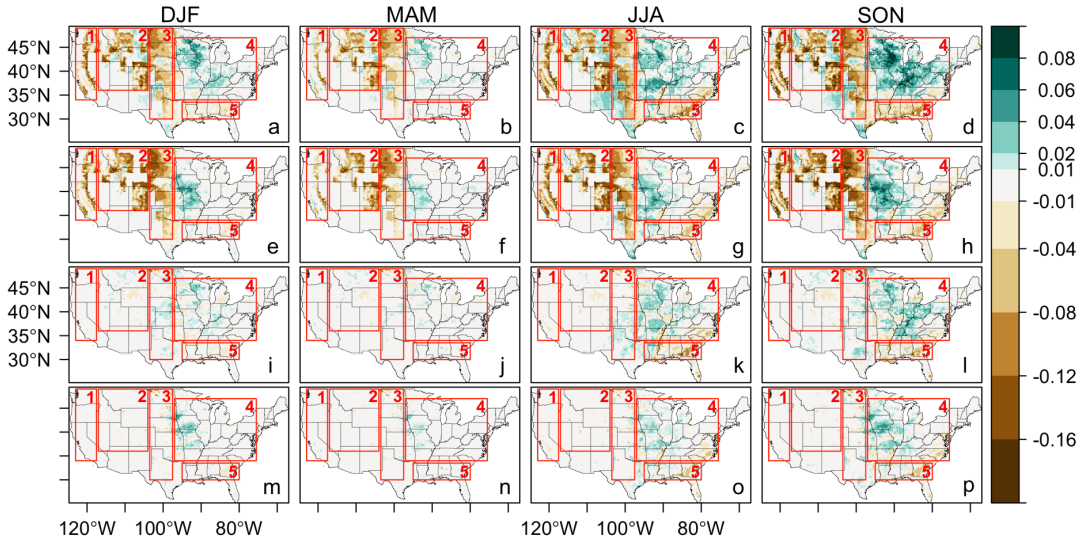

Figure 2.13: Seasonally averaged soil moisture difference (m3 m-3) at 40-100 cm soil depth (Perennial100_E1 - Control_E1) over one decade (2000-2009) for (a) DJF, (b) MAM, (c) JJA, and (d) SON. (e)-(h) Same as (a)-(d) but for difference of Perennial100_E8 minus Control_E8. (i)-(l) Same as (a)-(d) but for difference of Perennial25_E1 minus Control_E1. (m)-(p) Same as (a)-(d) but for difference of Perennial25_E8 minus Control_E8. Red rectangles outline five sub-regions for time series calculations.

Although time-averaged changes in soil moisture raise concerns associated with water depletion within the soil column, time series analyses of soil moisture provide insight into the progressive trend of these effects. Time series of soil moisture differences show seasonal and annual trends of soil moisture depletion, most notably at deeper soil depths (40-100 cm), with statistically significant decreasing trends in regions 2 and 3 under the full deployment scenario (Fig. 2.14, Table 2.3). In terms of seasonal differences, soil moisture associated with perennial bioenergy crops decreases during the growing season and then partially recharges from November until the following April over regions 1, 2, 3, and 5 (Fig.2.14a-c, e). This evolution of soil moisture differences is inversely related to changes in latent heat flux (for regions 1, 2, and 3) and is partially coincident with large-scale rainfall reduction (for region 5, see Figs. S2-S3). Under full bioenergy crop deployment, these differences are most noticeable with decreased soil moisture reaching 0.12 m3 m-3 over regions 2 and 3. Over the simulated decade and for these regions (western mountains and western Great Plains), soil moisture is depleted by roughly one-third of the initial soil moisture availability.

Moreover, soil moisture decreases progressively with each subsequent year for regions 2 and 3 under the full deployment scenario (with family-wise Type I error rate < 0.05 for simultaneous testing of all soil moisture difference trends; see Table 2.3). These progressive drying trends, however, are not evident in regions 1 (Pacific Coast), 4 (central/eastern U.S.), and 5 (Gulf Coast). Modified Mann-Kendall and sieve bootstrap tests show agreement in the trend test results.

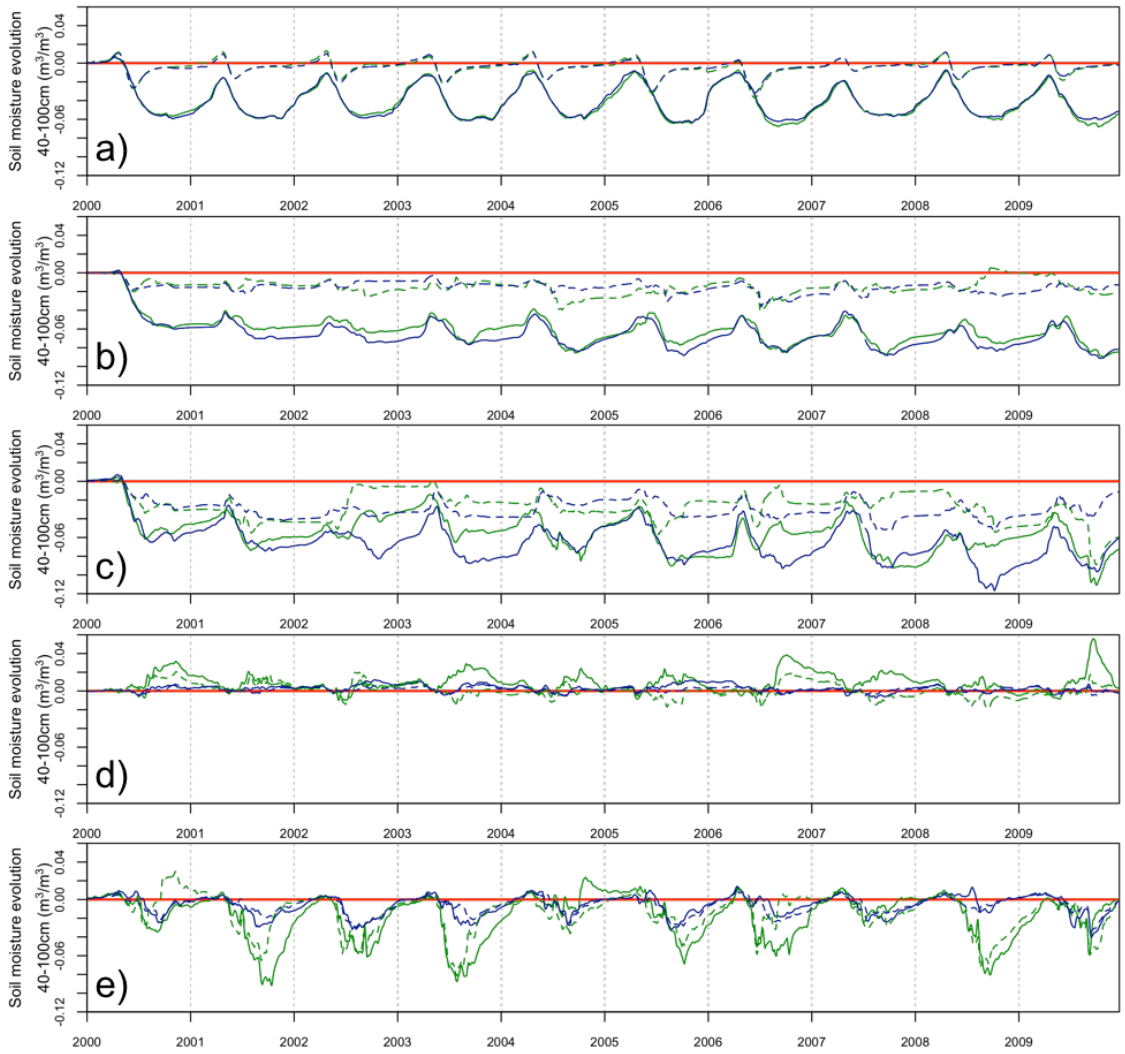

Figure 2.14: Spatially averaged soil moisture difference (m3 m-3) at 40cm-1m soil depth for grid cells undergoing land surface perturbation: (a) region 1, (b) region 2, (c) region 3, (d) region 4, and (e) region 5. Dark green and dark blue curves indicate ensemble member E1 and E8, respectively. Solid and dashed curves represent impact under Perennial100 scenario and Perennial25 scenario, respectively.

Table 2.3: Relative changes of soil moisture at the end of the 10^{th} simulation year (Perennial minus Control), normalized by the corresponding initial soil moisture at shallow and deeper soil depths.* and $^{\#}$ indicates statistically significant monotonic trends with 95% family-wise confidence (p-value < 0.001 for each test under the Bonferroni correction for multiple hypothesis tests), based on the Mann-Kendall test for serially correlated measurements and sieve bootstrap for Mann-Kendall tests, respectively.

Soil depth	Region	Perennial 100_E1 - Control_E1	Perennial 100_E8 - Control_E8	Perennial 25_E1 - Control_E1	Perennial 25_E8 - Control_E8
10-40 cm	1	-0.0997	-0.0863	0.0046	-0.0145
	2	-0.1469	-0.1223	-0.0364	-0.0137
	3	-0.1553	-0.1568	-0.1719	-0.0399
	4	0.0297	0.0159	0.0153	0.0124
	5	0.0007	-0.0022	-0.0075	-0.0037
40-100 cm	1	-0.2186	-0.2069	-0.0005	-0.0098
	2	-0.3483$^{\star,\#}$	-0.3353$^{\star,\#}$	-0.1001	-0.0586
	3	-0.3652$^{\star,\#}$	-0.3058$^{\star,\#}$	-0.3057	-0.0580
	4	0.0107	-0.0076	-0.0011	-0.0065
	5	0.0009	-0.0072	-0.0162	-0.0074

* and $^{\#}$ indicates statistically significant monotonic trends with 95% family-wise confidence (p-value < 0.001 for each test under the Bonferroni correction for multiple hypothesis tests), based on the Mann-Kendall test for serially correlated measurements and sieve bootstrap for Mann-Kendall tests, respectively.

2.3.2.4 Radiation balance

Changes in net radiation balance play an important role in driving the aforementioned hydroclimatic impacts. Overall, net radiation decreased, with the largest reduction occurring during summer (Fig.2.15a-d). These changes are largely responsible for the previously discussed changes in temperature and sensible heat flux. Under

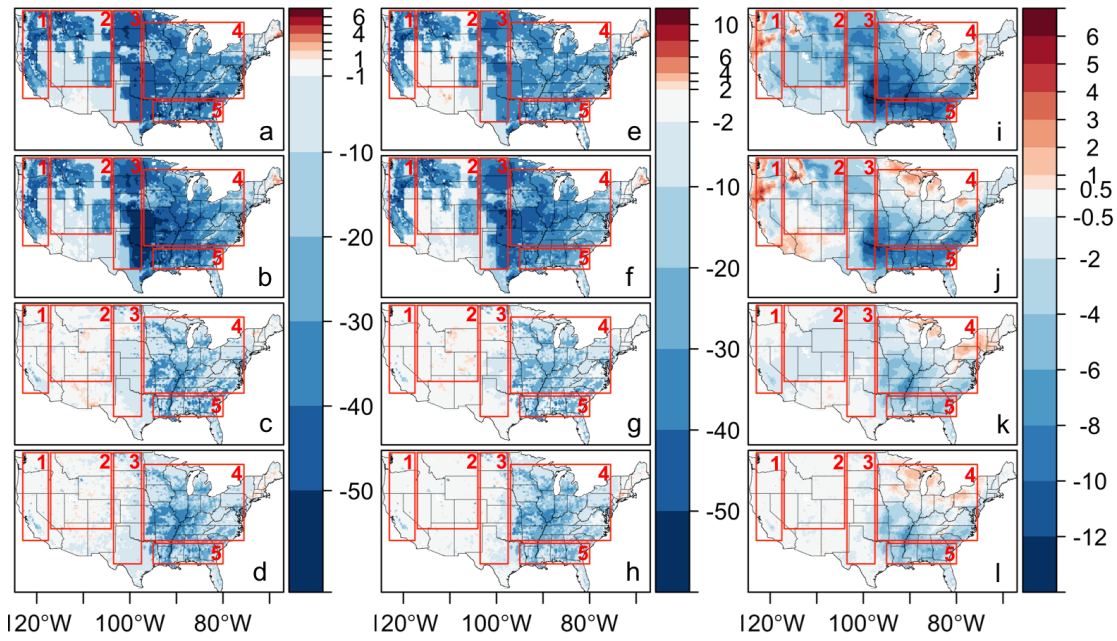

Figure 2.15: Summer (JJA) averaged net radiation difference (W m-2) over one decade (2000-2009) (a) Perennial100_E1 - Control_E1, (b) Perennial100_E8 - Control_E8, (c) Perennial25_E1 - Control_E1, and (d) Perennial25_E8 - Control_E8. (e)-(h) Same as (a)-(d) but for net shortwave radiation (W m-2). (i)-(l) Same as (a)-(d) but for net longwave radiation (W m-2).

the full deployment scenario, the largest reduction in net radiation (up to 60 W m-2) occurs over the southern Great Plains (mainly within region 3). Under the reduced deployment scenario, net radiation decreases 20-30 W m-2, primarily over the central/eastern U.S. (region 4) and Gulf Coast (region 5). According to time series plots of spatially-averaged net radiation differences, these decreases mainly occurred from mid-April to mid-October (Fig. S4).

The large-scale net radiation reduction is dominated by the decrease of shortwave radiation at the surface (Fig. 2.15e-h), resulting from enhanced surface reflectivity (Fig. 2.3c and 2.3g). Summer net shortwave decreases up to 50 W m-2 over the

southern Great Plains (the same region with maximum net radiation depletion), whereas the reduction of summer net longwave radiation peaks at roughly 12 W m-2 over southeastern areas of the U.S.

2.4 Discussion and Conclusions

Here we investigate hydroclimatic impacts of perennial bioenergy crop expansion over CONUS using continuous ensemble-based WRF simulations (2000 through 2009) and a suite of realistic deployment scenarios. Our results demonstrate that converting abandoned and degraded farmlands to perennial bioenergy croplands can lead to significantly cooler temperatures and potentially unintended consequences of soil moisture depletion for some regions of the U.S. Temperature decreases associated with perennial bioenergy crop deployment are largest over the Great Plains, generally 4-5 °C lower during the growing season compared to the unperturbed landscape (Figs. 2.7-2.8). Simulated soil moisture associated with perennial bioenergy crops shows a progressive decrease for some regions, most notably at deeper soil depths (40-100 cm). This decrease is most apparent under the full deployment scenario over the western Plains, with soil moisture depleted by 35% relative to the initial soil moisture availability (see Figs. 2.13-2.14). However, we note that, in general, smaller differences were evident under the reduced deployment scenario, although even in such instances soil moisture reduction was apparent (e.g., region 3; Table 2.3). Therefore, large-scale perennial bioenergy crop expansion over abandoned farmlands could have undesirable regional hydroclimatic consequences, but these effects are reduced for most areas undergoing small-scale deployment.

Biophysical parameters, including albedo, vegetation fraction and LAI, were shown

to serve as key factors characterizing hydroclimatic impacts due to perennial bioenergy crop expansion, in agreement with previous work focused on hypothetical landscape transitions (Georgescu *et al.* 2011; Davin *et al.* 2014; Zhu *et al.* 2017). Unlike previous studies (Le *et al.* 2011; Khanal *et al.* 2013; Abraha *et al.* 2015), changes in latent heat flux associated with perennial bioenergy crop expansion varied spatially (i.e., increased latent heat fluxes over some regions but minimal changes, or even decreases, over other regions).

We posit that a lack of statistically significant monotonic trends in soil moisture (Fig. 2.14; Table 2.3) accompanied by areas of regional cooling can be a determining factor in identifying suitable hotspots of bioenergy crop deployment. Perennial bioenergy crop expansion, therefore, could be sustainable in regions 4 and 5 (central/eastern U.S. and Gulf Coast states) based on the amount of soil moisture available during the annual cycle and the minimal to positive soil moisture changes simulated over the decadal timescale examined. Moreover, sections of Wisconsin and Missouri, extending eastward through the Ohio River Valley, could be posited as favorable locations for deployment due to seasonal soil moisture recharge (Fig. 2.13). Our results indicate statistically significant decreasing trends in soil moisture (up to 35% of initial soil moisture content) for regions 2 and 3 over the 10-year simulation period (see Table 2.3), highlighting these areas as potentially unsuitable. Although we do not observe a statistically significant trend in soil moisture for region 1 (i.e., California) the incomplete recovery of differences relative to the Control scenario during the winter season, does raise water resource concerns vis-à-vis depletion/interaction with the water table, which requires further investigation. However, it is worth noting that benefits may still exist as a decrease in runoff would lead to less soil erosion and therefore, could improve water quality over potential unsuitable areas.

We characterize the simulated large-scale hydroclimatic impacts associated with perennial bioenergy crop expansion as robust since the two sets of experiments (i.e., E8 and E1) converged to similar conclusions. Over most perennial bioenergy crop deployment regions, the best (i.e., E8) and least (i.e., E1) skilled ensemble members yielded similar results in terms of the magnitude and extent of regional cooling, changes in latent and sensible heat fluxes, and soil moisture impacts. Additionally, the overlaid climate variability ranges associated with the mean annual cycle of sub-regionally averaged cooling and changes of surface energy balance components between the aforementioned two ensemble members provide further confidence in our results. It is important to mention that the predicted temperature in our simulations exhibits reduced scattering compared to precipitation. This suggests that the errors observed in precipitation, owing to utility of different cloud microphysics parameterizations, do not have significant impact on the dynamics simulated by WRF. If these errors were important, they would have affected the dynamics through temperature changes caused by the release or absorption of latent heat. Consequently, the scattering in temperature and precipitation would have been closely correlated. However, this was not observed in our simulations, consistent with previous research (e.g., Done *et al.* 2005; Okalebo *et al.* 2016). Nevertheless, from a purely physics and model development perspective additional insights characterizing the parameterization aspects leading to quantitative determination detailing differences in simulated results (e.g., what particular aspects of parameterized features contributes to this variability) is an important research avenue for pursuit, but is beyond the focus of this manuscript.

Finally, the principal highlights of this research establish a framework of feedback assessment between LULCC and water resource impacts where analogous energy pathways involving landscape modification are being considered (e.g., natural landscape

conversion to oil palm in Indonesia). Via identification of suitable hotspots of bioenergy crop deployment, due to simultaneous regional-scale cooling in conjunction with minimal adverse effects on soil moisture, we also identify areas wherein cultivation can effectively reduce projected warming due to large-scale climate change.

Chapter 3

SPATIO-TEMPORAL MODELING FOR REGIONAL CLIMATE MODEL COMPARISON: APPLICATION ON PERENNIAL BIOENERGY CROP IMPACTS

3.1 Introduction

Deployment of perennial bioenergy crops is an alternative energy pathway to mitigate climate change, increase energy independence, stabilize energy prices, and achieve hydroclimatic sustainability in some marginal lands. Previous studies used regional climate models (RCMs) to quantify perennial bioenergy crops impacts (Anderson *et al.* 2013; Georgescu *et al.* 2011; Khanal *et al.* 2013; Wagner *et al.* 2017; Wang *et al.* 2017). However, RCMs with different physics parameterizations could generate significantly different outputs, leading to uncertainties of results to be examined. Therefore, in order to examine the robustness of simulated perennial bioenergy crop impacts, it is essential to assess the significance of factors associated with RCM performance.

The uncertainties of RCM outputs have been studied using both descriptive and inferential statistics. Specifically, Taylor diagrams and Hövmoller diagrams have been applied to evaluate RCM simulation skill using multiple performance metrics (Hovmöller 1949; Taylor 2001; Wang *et al.* 2017). However, the abovementioned diagrams cannot be used to assess the significance of factors associated with simulation skill. Sansom *et al.* (2013) assigned different weights to ensemble members of RCMs, based on an ANOVA framework. This method did not take into account spatiotemporal dependencies. Kang *et al.* (2012) applied hierarchical Bayesian spatial random-effects model to quantify the climate signal of individual RCMs. Although spatially correlated

processes could be captured, the proposed framework did not include a temporal component. Given that spatio-temporal dependencies are inherent to RCM outputs, spatio-temporal statistical models are needed.

There is a variety of Bayesian hierarchical spatio-temporal models (BHM) and corresponding R packages, such as spBayes, spTDyn, spate, spTimer (Finley *et al.* 2013; Bakar *et al.* 2016; Sigrist *et al.* 2015; Bakar *et al.* 2016). For spatio-temporal modeling of lattice data, the R package named surveillance can be used (Meyer *et al.* 2014); nlme and lme4 can model spatial and temporal effects by fitting linear mixed models (Pinheiro 2009; Bates *et al.* 2014); CARBayesST can implement hierachical spatio-temporal generalized linear mixed models (Lee *et al.* 2017).

Alternative methods have been used for estimating the posterior distribution of BHM, such as Markov chain Monte Carlo (MCMC) sampling (Christian and Casella 1999), and Integrated nested Laplace approximations (INLA) (Rue *et al.* 2009; Blangiardo and Cameletti 2015). MCMC methods refers to sampling from a probability distribution by constructing a Markov chain that has the desired posterior distribution as its stationary distribution. Several MCMC algorithms have been used to approximate multidimensional integrals. For example, Metropolis–Hastings algorithm generates random samples using a proposal density with some probability of acceptance and rejection (Metropolis *et al.* 1953; Hastings 1970); and Gibbs samplig, which samples from the conditional posterior distributions exactly (hence, it does not require any 'tuning') (Geman and Geman 1987). However, MCMC is not straightforward to implement and may be slow to converge. INLA overcomes these issues as posteriors are estimated using numerical approximations (hence, random sampling is avoided). It is considered to be superior to MCMC in terms of accuracy and

computational efficiency, although it is restricted for only analyzing latent Gaussian models.

To our knowledge, a limited number of research works compare alternative parameter estimation approaches and models with different spatio-temporal autocorrelation structures. In this study, multiple spatio-temporal models are compared for modeling RCM outputs; the motivating application aims to evaluate perennial bioenergy crop impacts. More specifically, the work in this chapter investigates the following research questions:

a. Do physics parameterizations and observations have a significant impact on WRF control simulations?

b. Is WRF-simulated temperature impact associated with perennial bioenergy crops robust to alternative physics parameterizations?

c. Which spatio-temporal residual correlation structure is the most appropriate given the fixed effects?

This chapter is arranged as follows. Section 3.2 presents a review of three commonly used Bayesian hierarchical spatio-temporal models for lattice data, as well as the methodology of modeling RCM output ensembles. The application is presented and discussed in Section 3.3. Concluding remarks and suggestions for future work are discussed in Section 3.4.

3.2 Methodology

3.2.1 Bayesian hierarchical spatio-temporal models for lattice data

Let $A = \{A_1, \ldots, A_S\}$ be a set of S non-overlapping lattice units. Data are collected with S spatial units and T consecutive time periods, available in a $S \times T$ rectangular array. $\boldsymbol{Y} = (\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_T}) = (Y_{11}, \ldots, Y_{S1}, \ldots, Y_{1T}, \ldots, Y_{ST})$ denotes the vector of continuous response variable. For one lattice unit s at time period t, where $s = 1, \ldots, S$ and $t = 1, \ldots, T$, $\boldsymbol{x_{st}} = \big(x_{st1}, \ldots, x_{stp}\big)$ is a vector of p known covariates.

Bayesian hierarchical space-time models for lattice data can be considered within the linear mixed model framework (Lee *et al.* 2017); the general formulation is given by:

$$\begin{aligned} Y_{st}|\mu_{st} &\sim N(\mu_{st}, \sigma^2), \\ \mu_{st} &= \boldsymbol{x}_{\boldsymbol{st}}^{\mathsf{T}}\boldsymbol{\beta} + M_{st}, \\ \boldsymbol{\beta} &\sim N(\boldsymbol{\mu_\beta}, \boldsymbol{\Sigma_\beta}). \end{aligned} \tag{3.1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ is a vector of covariate parameters, a multivariate Gaussian prior with mean $\boldsymbol{\mu_\beta}$ and diagonal variance matrix $\boldsymbol{\Sigma_\beta}$. The M_{st} term is a latent component that captures remaining spatio-temporal autocorrelation for lattice unit s at time period t.

3.2.2 Spatio-temporal models for M_{st}

The spatio-temporal autocorrelation of data, $\boldsymbol{M}$, where $\boldsymbol{M} = (\boldsymbol{M_1}, \ldots, \boldsymbol{M_T})$ and $\boldsymbol{M_t} = (M_{1t}, \ldots, M_{St})$, are commonly modeled as random effects by the class of

Bayesian Conditional Autoregressive (CAR) prior distributions. CAR distributions are a type of Markov random field model, meaning that adjacent variables (i.e., in space or time) are autocorrelated, whereas variables for non-neighboring lattice units are conditionally independent given the remaining variables. The autocorrelation with respect to space is determined by a $S \times S$ spatial neighborhood matrix $\boldsymbol{W} = (w_{sj}), j = 1, \ldots, S$. $\boldsymbol{W}$ is symmetric, consisting of binary elements. (w_{sj}) has value 1 if lattice units (A_s, A_j) are close in space (i.e., share a common border) and is zero otherwise. Additionally, $(w_{jj}) = 0$. Similarly, the binary $T \times T$ temporal neighborhood matrix is defined as $\boldsymbol{D} = (d_{tj}), j = 1, \ldots, T$, where $d_{tj} = 1$ if $|t - j| = 1$ and $d_{tj} = 0$ otherwise. Based on the neighboring information, three spatio-temporal autocorrelation structures $\boldsymbol{M}$ are considered in this chapter, namely STCARlinear, STCARanova, and STCARar. The implementation of these three models can be found in R package CARBayesST (Lee *et al.* 2017).

3.2.2.1 Model-1: STCARlinear

STCARlinear is a spatially varying linear time trends model (Bernardinelli *et al.* 1995). It is formulated as:

$$M_{st} = \beta_1 + \phi_s + (\alpha + \delta_s)t^*, \tag{3.2}$$

where $\beta_1 + \phi_s$ denotes region-specific intercept and $\alpha + \delta_s$ represents region-specific temporal trend. $t^* = \dfrac{t - \bar{t}}{T} = \dfrac{t - \sum_{t=1}^{T} t/T}{T}$ denotes the linear temporal covariate running over a centered unit interval. Specifically, β_1 is the intercept in $\boldsymbol{\beta}$ (Eq. (3.1)) and α denotes a global slope parameter which is normally distributed with mean μ_α and variance σ_α^2.

The spatial correlation in ϕ_s and δ_s is enforced by the CAR prior (Leroux *et al.*

2000). These parameters are mean centered (Lee *et al.* 2017) and their conditional distributions are formulated as::

$$\begin{aligned}
\phi_s|\boldsymbol{\phi_{-s}}, \boldsymbol{W} &\sim N\left(\frac{\rho_{int}\sum_{j=1}^{S} w_{sj}\phi_j}{\rho_{int}\sum_{j=1}^{S} w_{sj} + 1 - \rho_{int}}, \frac{\tau_{int}^2}{\rho_{int}\sum_{j=1}^{S} w_{sj} + 1 - \rho_{int}}\right), \\
\delta_s|\boldsymbol{\delta_{-s}}, \boldsymbol{W} &\sim N\left(\frac{\rho_{slo}\sum_{j=1}^{S} w_{sj}\delta_j}{\rho_{slo}\sum_{j=1}^{S} w_{sj} + 1 - \rho_{slo}}, \frac{\tau_{slo}^2}{\rho_{slo}\sum_{j=1}^{S} w_{sj} + 1 - \rho_{slo}}\right),
\end{aligned} \tag{3.3}$$

where ρ_{int} and ρ_{slo} are spatial dependence parameters which are assigned Uniform$(0, 1)$ priors; values of 1 (i.e., intrinsic CAR prior proposed by Besag *et al.* (1991)) and 0 correspond to spatial dependence and independence, respectively. In addition, τ_{int}^2 and τ_{slo}^2 are assigned inverse-gamma priors with shape a and scale b. In this chapter, the corresponding hyper priors $(a, b, \mu_\alpha, \sigma_\alpha^2)$ are specified to be 0.001, 0.001, 0, and 1000, respectively [in accordence with Lee *et al.* (2017) so that the priors are weakly informative]. Using the above structure, spatio-temporal autocorrelation of data is a function of time with spatial dependencies assumed by their priors; the spatial associations are related to the spatial weight matrix $\boldsymbol{W}$.

3.2.2.2 Model-2: STCARanova

STCARanova consists of three components of spatio-temporal variation: an overall spatial effect $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_S)$, an overall temporal trend $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_S)$, and independent space-time interactions $\boldsymbol{\gamma} = (\gamma_{11}, \ldots, \gamma_{ST})$ (Knorr-Held and Besag 1998; Knorr-Held 1999):

$$M_{st} = \phi_s + \delta_t + \gamma_{st}. \tag{3.4}$$

The first (two, mean centered) components are assigned the CAR prior proposed by Leroux *et al.* (2000):

$$\begin{aligned}
\phi_s|\boldsymbol{\phi_{-s}}, \boldsymbol{W} &\sim N\left(\frac{\rho_S \sum_{j=1}^{S} w_{sj}\phi_j}{\rho_S \sum_{j=1}^{S} w_{sj} + 1 - \rho_S}, \frac{\tau_S^2}{\rho_S \sum_{j=1}^{S} w_{sj} + 1 - \rho_S}\right), \\
\delta_t|\boldsymbol{\delta_{-t}}, \boldsymbol{D} &\sim N\left(\frac{\rho_T \sum_{j=1}^{T} d_{tj}\delta_j}{\rho_T \sum_{j=1}^{S} d_{tj} + 1 - \rho_T}, \frac{\tau_T^2}{\rho_T \sum_{j=1}^{S} d_{tj} + 1 - \rho_T}\right),
\end{aligned} \tag{3.5}$$

Optional set of spatial and temporal interaction effects $\boldsymbol{\gamma_{st}}$ can be specified in the model to capture nonlinear patterns. Four possible types of interaction assume: independence, purely spatial dependence, purely temporal dependence, and spatio-temporal autocorrelation (Knorr-Held 1999). In this chapter, however, the space-time interaction is not included in the model. The avoidance of interaction term could reduce complexity while increasing flexibility for space-time modeling (López-Quilez and Munoz 2009).

In Eq.(3.5), ρ_s and ρ_t are spatially and temporally dependent parameters, respectively, distributed with Uniform(0.1) priors. Similarly, τ_s^2 and τ_t^2 are spatially and temporally dependent parameters, respectively, which are assigned inverse-gamma(a, b) priors. The hyper parameters (a, b) are both chosen to be 0.001 [in accordance with the suggestion of Lee *et al.* (2017) as weakly informative priors].

3.2.2.3 Model-3: STCARar

STCARar models the spatio-temporal structure as a multivariate first order autoregressive process with a spatially correlated precision matrix (Rushworth *et al.*

2014). The model specification is given by:

$$
\begin{aligned}
&M_{st} = \phi_{st}, \\
&\boldsymbol{\phi_1} = (\phi_{11}, \ldots, \phi_{S1}) \sim N(\mathbf{0}, \tau^2 Q(\boldsymbol{W}, \rho_S))^{-1}) \\
&\boldsymbol{\phi_t} | \boldsymbol{\phi_{t-1}} \sim N(\rho_T \boldsymbol{\phi_{t-1}}, \tau^2 Q(\boldsymbol{W}, \rho_S))^{-1}, t = 2, \ldots, T,
\end{aligned}
\tag{3.6}
$$

where $\boldsymbol{\phi_t}$ is the vector of random effects at time period t; ρ_T denotes a temporal autoregressive parameter and $\rho_T \boldsymbol{\phi_{t-1}}$ induces temporal autocorrelation; variance $\tau^2 Q(\boldsymbol{W}, \rho_S))^{-1}$ imposes spatial autocorrelation with:

$$
Q(\boldsymbol{W}, \rho_S) = \rho_S \left[\text{diag}(\boldsymbol{W}\mathbf{1}) - \boldsymbol{W} \right] + (1 - \rho_S)\boldsymbol{I}, \tag{3.7}
$$

where $\mathbf{1}$ is the $S \times 1$ vector of ones and $\boldsymbol{I}$ is the $S \times S$ identity matrix. The precision matrix $Q(\boldsymbol{W}, \rho_s)$ corresponds to the CAR prior (Leroux *et al.* 2000), with specification at time period 1:

$$
\phi_{s1} | \boldsymbol{\phi_{-s1}}, \rho_S, \tau^2, \boldsymbol{W} \sim N \left(\frac{\rho_S \sum_{j=1}^{S} w_{sj} \phi_{j1}}{\rho_S \sum_{j=1}^{S} w_{sj} + 1 - \rho_S}, \frac{\tau^2}{\rho_S \sum_{j=1}^{S} w_{sj} + 1 - \rho_S} \right). \tag{3.8}
$$

In Eq.(3.8), ρ_S controls the level of spatial smoothness, with $\rho_S = 1$ leading to CAR model proposed by Besag *et al.* (1991) while $\rho_S = 0$ inducing identical and independent (iid) normal prior distributions. In this Chapter of study, ρ_S is fixed at unity.

3.3 Spatio-temporal modeling for simulated temperature differences

3.3.1 Description of datasets

The analyzed data are seasonally averaged WRF-simulated temperatures from 2000 to 2009 over the conterminous U.S. Two types of datasets are analyzed (Table 3.1). The first type relates to simulation bias, i.e., the difference of reproduced temperature

and the corresponding observations. Sixteen scenarios are included in this group: scenarios vary by choices of microphysics schemes, cumulus schemes, utility of spectral nudging, and observations. Each of the aforementioned factors includes two levels. The second type of dataset relates to biofuel impact: the difference of reproduced temperature and temperature under full-deployment scenario of perennial bioenergy crops expansion. In the second data type, only two scenarios are included, varied by combinations of physics parameterizations for best and worst skilled model, selected based on the results presented in Chapter 2. Details with regard to the experimental design can be found in Wang *et al.* (2017). Both types of datasets are gridded data with spatio-temporal dependence. In this chapter, the datasets were resampled using bilinear interpolation to include $S = 348$ pixels at each time period, with $T = 40$ time periods in total (seasonal values in consecutive 10 years)[1].

3.3.2 Spatio-temporal statistical modeling

For each scenario, the expected response (i.e., differences of reproduced temperature and the corresponding observations for first type of datasets; and temperatures changes associated with perennial bioenergy crop expansion) was modeled using the following specification:

$$\mu_{st} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + M_{st} \tag{3.9}$$

where x_1, x_2, x_3 are fixed effects of seasons (i.e., the effects of Spring, Summer, and Fall relative to Winter, respectively), while β_1, β_2, and β_3 denote the coefficients of corresponding factors. M_{st} represents the spatio-temporal random effect, which may

[1]Bilinear interpolation is a resampling method which estimates a new pixel value by a weighted average of the four nearest pixels, according to distance.

Table 3.1: Description of datasets.

Type of dataset	Scenarios	Microphysics	Cumulus physics	Spectral nudging technique	Observed data
simulation bias	S1	WSM3	Kain–Fritsch	No	DW
	S2	WSM3	Kain–Fritsch	Yes	DW
	S3	WSM3	Grell 3D	No	DW
	S4	WSM3	Grell 3D	Yes	DW
	S5	WDM6	Kain–Fritsch	No	DW
	S6	WDM6	Kain–Fritsch	Yes	DW
	S7	WDM6	Grell 3D	No	DW
	S8	WDM6	Grell 3D	Yes	DW
	S1	WSM3	Kain–Fritsch	No	GC
	S2	WSM3	Kain–Fritsch	Yes	GC
	S3	WSM3	Grell 3D	No	GC
	S4	WSM3	Grell 3D	Yes	GC
	S5	WDM6	Kain–Fritsch	No	GC
	S6	WDM6	Kain–Fritsch	Yes	GC
	S7	WDM6	Grell 3D	No	GC
	S8	WDM6	Grell 3D	Yes	GC
biofuel impact	S1	WSM3	Kain–Fritsch	No	N/A
	S8	WDM6	Grell 3D	Yes	N/A

follow the specifications from equations 3.2 to 3.6, i.e., STCARlinear, STCARanova, STCARar.

Posterior samples of spatial and temporal dependence parameters for each model and each scenario were compared using box plots. The model that had consistent spatio-temporal structure across scenarios was applied on the pooled data (combined scenarios), in order to assess the significance of factors associated with RCM performance and the robustness of estimated perennial bioenergy crop impacts.

3.3.3 Modeling multiple spatio-temporal processes simultaneously

When modeling multiple spatio-temporal processes, the aforementioned specifications cannot be applied, since they are designed for a single spatio-temporal process. Here, we propose a new method to model several spatio-temporal processes simultaneously. First of all, each univariate spatial-temporal process is modeled individually using each candidate model. Then the performances of each model can be evaluated: the preferred candidate model should be the one that best captures the spatio-temporal autocorrelation of each single spatio-temporal process. The spatio-temporal autocorrelation structure of the selected model can then be considered as a consensus structure for all univariate processes. Lastly, multiple spatio-temporal processes are pooled together and modeled, using the selected specification. The neighborhood matrix $\boldsymbol{W} = (w_{sj})$ is modified to be $\boldsymbol{W}_{consensus} = \boldsymbol{I} \bigotimes \boldsymbol{W}$, where $\boldsymbol{I}$ is the m by m identity matrix, with m denoting the number of spatio-temporal processes.

It was assumed that all scenarios of the same data type have a consistent spatio-temporal structure: this structure was selected based on the results of scenario-specific models. Scenario-combined models were different from scenario-specific models only with regard to the fixed effects part. For simulation bias, the mean of the set of processes were modeled using the specification as Eq. (3.1), where $\boldsymbol{X}$ contains columns (x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_1x_4, x_1x_5,x_1x_6, x_1x_7, x_2x_4, x_2x_5, x_2x_6, x_2x_7, x_3x_4, x_3x_5, x_3x_6, x_3x_7), with (x_1, x_2, x_3) denoting fixed effects for seasons , (x_4, x_5, x_6, x_7) denoting factors for microphysics parameterizations, cumulus scheme parameterizations, spectral nudging and observations, respectively, and x_1x_4, x_1x_5,x_1x_6, x_1x_7, x_2x_4, x_2x_5, x_2x_6, x_2x_7, x_3x_4, x_3x_5, x_3x_6, x_3x_7) denoting interactions. $\boldsymbol{\beta}$ are the corresponding coefficients.

For biofuel impact, datasets were modeled using the the same specification as Eq.(3.1), but now $\boldsymbol{X}$ contains columns $(x_1, x_2, x_3, x_4, x_1x_4, x_2x_4, x_3x_4)$, where (x_1, x_2, x_3) represent fixed (seasonal) effects, x_4 denotes alternative physics characterizations, and (x_1x_4, x_2x_4, x_3x_4) represent the corresponding interaction effects.

3.3.4 Model selection and the selection criteria

After the spatio-temporal autocorrelation structure have been chosen, model selection was conducted in order to derive a parsimonious model. The final models were selected by evaluating criteria which favor high likelihood while penalizing model complexity. Three criteria are frequently used: Akaike Information Criterion, Bayesian Information Criterion, and Deviance Information Criterion (i.e., AIC, BIC, and DIC, respectively). Let k be the number of estimated parameters in the model for a particular data; let $\hat{L}$ be the maximized value of the likelihood function for the model and n the sample size. AIC is formulated as (Akaike 1998):

$$AIC = -2\ln(\hat{L}) + 2k \tag{3.10}$$

Compared to AIC, BIC contains a modified penalty term for the number of parameters (Schwarz *et al.* 1978):

$$BIC = -2\ln(\hat{L}) + \ln(n)k \tag{3.11}$$

In bayesian analysis, DIC is frequently used; it is based on posterior distributions of the model by MCMC simulation, and follows similar ideas to AIC and BIC, as it evaluates both "goodness of fit" and "complexity" of the model. More specifically, fit is measured via the deviance as $D(\theta) = -2log(p(\text{data}|\theta))$, where θ are the unknown parameters of the model and $p(\text{data}|\theta)$ is the likelihood function. Complexity is measured via either $p_D = E(D) - D(\bar{D})$ (i.e., posterior mean deviance minus deviance

evaluated at the posterior mean of the parameters) (Spiegelhalter *et al.* 2002), or by $p_D = 1/2\text{var}(\hat{D})$ (Gelman *et al.* 2014). In general, smaller values of AIC, BIC, or DIC indicate superior specifications; models with value differences smaller than 5 in the above criteria could be considered as equivalent. In this case, combining these equivalent models via model averaging may result in a more robust specification (Hoeting *et al.* 1999).

In this chapter, we first applied the aforementioned Bayesian hierachical specifications for each scenario; simultaneous modeling of multiple spatio-temporal processes was based on the relative degree of consensus in space-time parameter estimation and DIC. For multiple spatio-temporal processes, we conducted conventional (frequentist) linear regression and applied exhaustive model selection based on AIC to exclude non-significant factors. Using this procedure, the number of model candidates for modeling multiple spatio-temporal processes was brought down to 5 models. BHM was performed for these specifications and DIC was used to select the final spatio-temporal model.

3.4 Results

3.4.1 Model comparison for a single scenario

For each scenario, a single chain with a total of 100,000 iterations was simulated. The number of iterations for the burnin period was 10,000, and the thinning rate was 100 so the number of samples used for the estimation of the parameters was 900. MCMC was evaluated using the Geweke diagnostic (Geweke *et al.* 1991), which should

be within the range of -2 and 2. The minimum acceptable effective sample size was set to be equal to 150.

Table 3.2 presents DIC values for all combinations of datasets, scenarios, and statistical models. For all scenarios of simulation bias data-type, STCARar achieves the lowest DIC consistently, while STCARlinear attains higher DIC than STCARanova in all scenarios except S5, S6, S8 and S14. On the contrary, STCARar models achieves the highest DIC and STCARlinear the lowest for biofuel impact data-type. Thus, STCARlinear fits the simulation bias data-type worse than the other two spatio-temporal structures, whereas it fits the biofuel impact data-type better than the other structures. STCARar, on the other hand, fits the simulation bias data-type better, but achieves the worst performance for biofuel impact data-type.

3.4.2 Spatio-temporal modeling of individual scenarios

3.4.2.1 Fixed effect estimates

The medians of fixed effects are very close across spatio-temporal models; on the other hand, the 95% confidence intervals (CIs) differ dramatically (Table 3.3). This finding is consistent across the 3 models. For example, the estimated medians of simulation biases based on scenario S1 are equal to 2.20, 0.46, 3.51, and 2.75 for the Intercept, Spring-Winter difference, Summer-Winter difference, and Fall-Winter difference (i.e., $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$), respectively. The widest 95% CIs, however, are observed using STCARar, whereas the narrowest using STCARlinear. It is worth noting that there are no significant differences for fixed effect estimates derived from alternative approximation methods (i.e., MCMC vs INLA) as the corresponding 95%

Table 3.2: DIC values for single scenarios. The scenario with the smallest values per scenario is highlighted in bold.

Type of dataset	Scenarios	STCARlinear	STCARanova	STCARar
simulation bias	S1	30556.8	30299.61	**27239.83**
	S2	29659.08	29469.15	**26759.66**
	S3	30323.23	30067.83	**26602.55**
	S4	29607.56	29502.47	**26718.72**
	S5	29185.82	28949.95	**26354.89**
	S6	27905.07	27948.96	**26548.55**
	S7	29028.95	28669.41	**26214.85**
	S8	28002.11	28040.67	**27057.49**
	S1	30610.51	30308.43	**27032.93**
	S2	29696.17	29525.09	**27066.27**
	S3	30288.61	30054.88	**26887.28**
	S4	29586.9	29494.1	**26848.38**
	S5	29120.41	28823.15	**26502.2**
	S6	27803.51	27841.85	**26549.3**
	S7	29221.08	28876.18	**26213.22**
	S8	27715.63	27701.42	**26467.08**
biofuel impact	S1	**39255.33**	39354.91	39368.12
	S8	**38915.38**	39041.54	39098.31

CIs are overlaid. An algorithm for STCARar using INLA is not currently available, so the corresponding results are not included in Table 3.3.

Results show that simulated temperature biases differ significantly by season. The simulated temperature relative to observation DW is approximately 2.2 °C lower in Winter. However, biases are 0.46, 3.5 and 2.7 °C higher in Spring, Summer, and Fall than in Winter, respectively (Table 3.3). Therefore, the simulated temperatures are 1.74 °C lower than observation DW, whereas 1.31, and 0.57 °C higher in Summer, and Fall, respectively.

Table 3.3: Fixed effect estimates of simulation bias data-type (scenario S1)

Fixed effect	Model	95% Confidence Interval								
		CARlinear			CARanova			CARar		
		Median	2.50%	97.50%	Median	2.50%	97.50%	Median	2.50%	97.50%
β_0	MCMC	-2.2	-2.26	-2.14	-2.21	-2.38	-2.05	-2.21	-2.49	-1.95
	INLA	-2.13	-2.21	-2.06	-2.21	-2.4	-2.01			
β_1	MCMC	0.46	0.38	0.54	0.47	0.2	0.72	0.47	-0.01	0.91
	INLA	0.46	0.37	0.54	0.46	0.19	0.73			
β_2	MCMC	3.51	3.44	3.59	3.53	3.27	3.82	3.52	3.09	3.99
	INLA	3.51	3.43	3.59	3.52	3.25	3.79			
β_3	MCMC	2.75	2.67	2.84	2.77	2.53	3.01	2.78	2.36	3.26
	INLA	2.75	2.67	2.84	2.77	2.5	3.04			

3.4.2.2 Spatio-temporal correlation

For STCARlinear, the range of spatial intercepts and slopes overlay consistently across scenarios (Fig.3.1). When modeling simulation bias, the means of spatially dependent intercepts lie between 0.9 and 0.95 whereas the majority of means of spatially dependent slopes range from 0.8 to 0.9 (Fig. 3.1(a) and (b)). The spatially dependent variances of slopes differ, with larger magnitudes across scenarios relative to the ones that correspond to intercepts (Fig.3.1(c) and (d)). The overall spatial intercepts are generally greater than 0 for all scenarios (Fig. 3.1(e)). For biofuel impacts, however, the differences of posterior samples across scenarios are relatively small (Fig. 3.1(f) - (j)).

For STCARanova, although the posterior distributions of spatially dependent means are similar across scenarios, the rest of posterior samples differ dramatically (Fig.3.2(a)-(e)). Specifically, temporally dependent variances and error variances for simulation bias data, possess significantly different posterior distributions across scenarios. However, posterior distributions related to biofuel impacts do not differ dramatically, similar to what is observed for STCARlinear.

Posterior samples of STCARar differ largely across scenarios for both simulation bias and biofuel impact data types (Fig. 3.3). Specifically, posterior samples of

temporally autoregressive parameters and variances of spatial autocorrelations could differ on the order of five to seven times across scenarios (Fig. 3.3(b) and (c)).

Taking into account the model comparisons presented above and the interest of combining scenarios in a single specification, STCARlinear appears to be the best choice: a consistent spatial-temporal structure can be assumed across scenarios using this structure, given that seasonal factors are included.

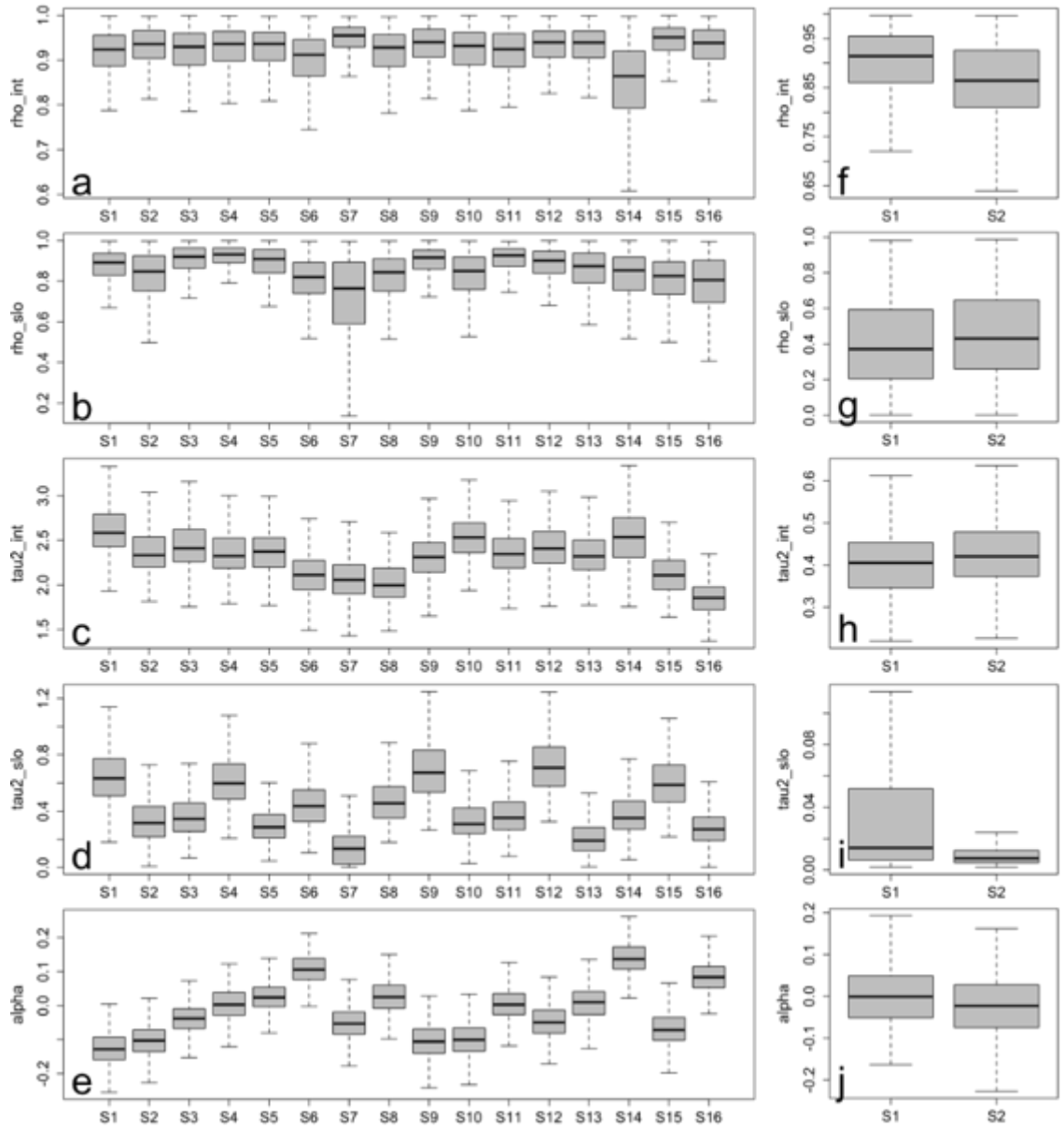

Figure 3.1: (a) Box-plots of posterior samples of spatio-temporal random effects using STCARlinear. Each box plot corresponds to a scenario-specific model: (a) mean of spatially dependent intercept, associated with simulation bias data-type; (b)-(e) the same as (a) but for mean of spatially dependent slope, variance of spatially dependent intercept, variance of spatially dependent slope, and overall slope parameter, respectively; (f)-(g) the same as (a)-(e), but associated with biofuel impact data-type.

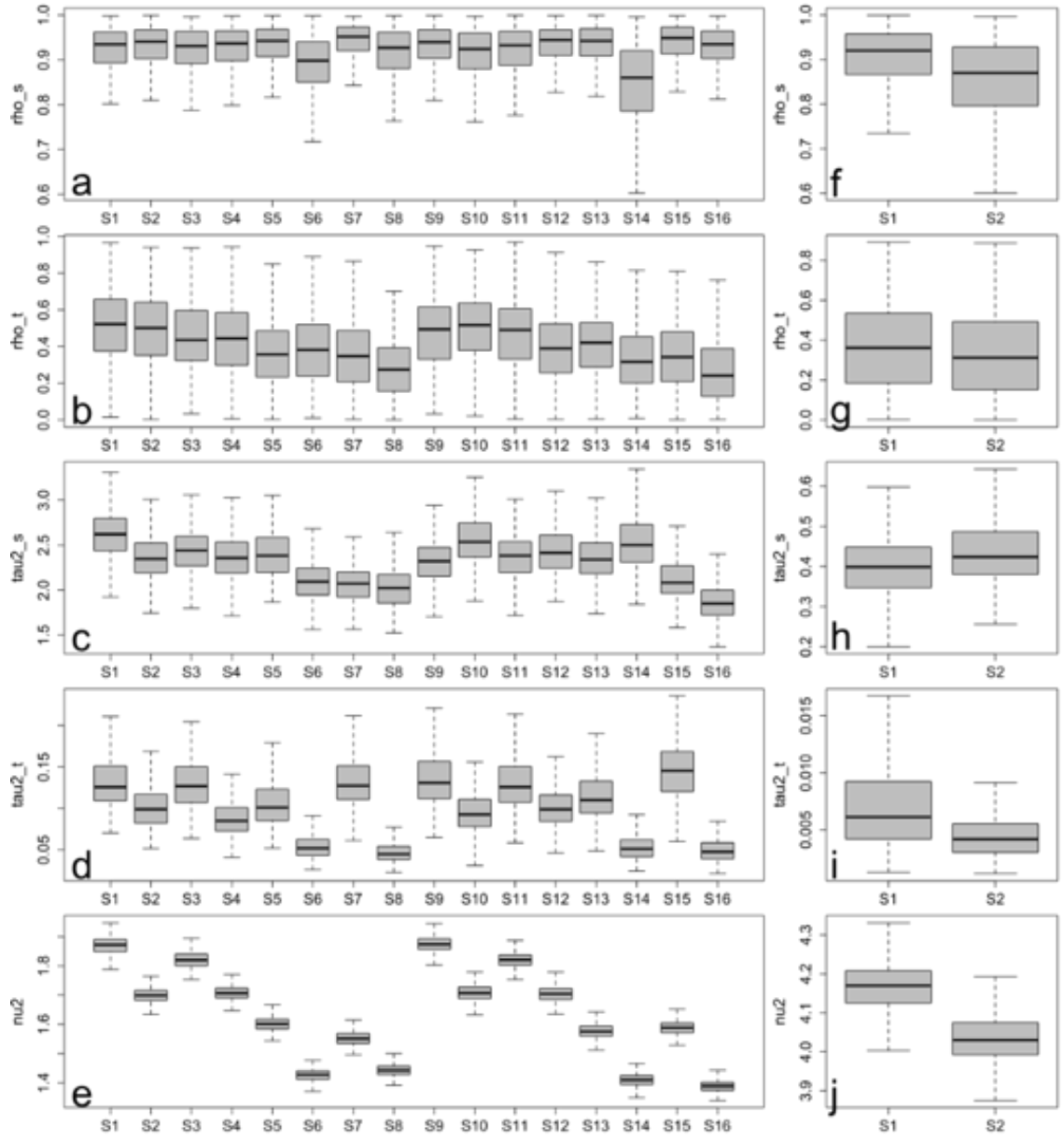

Figure 3.2: Box-plots of posterior samples of spatio-temporal random effects using STCARanova. Each box plot corresponds to a scenario-specific model: (a) spatially dependent mean, associated with simulation bias data-type; (b)-(e) the same as (a) but for temporally dependent mean, spatially dependent variance, temporally dependent variance, and prior for the Gaussian error variance, respectively; (f)-(g) the same as (a)-(e), but associated with biofuel impact data-type.

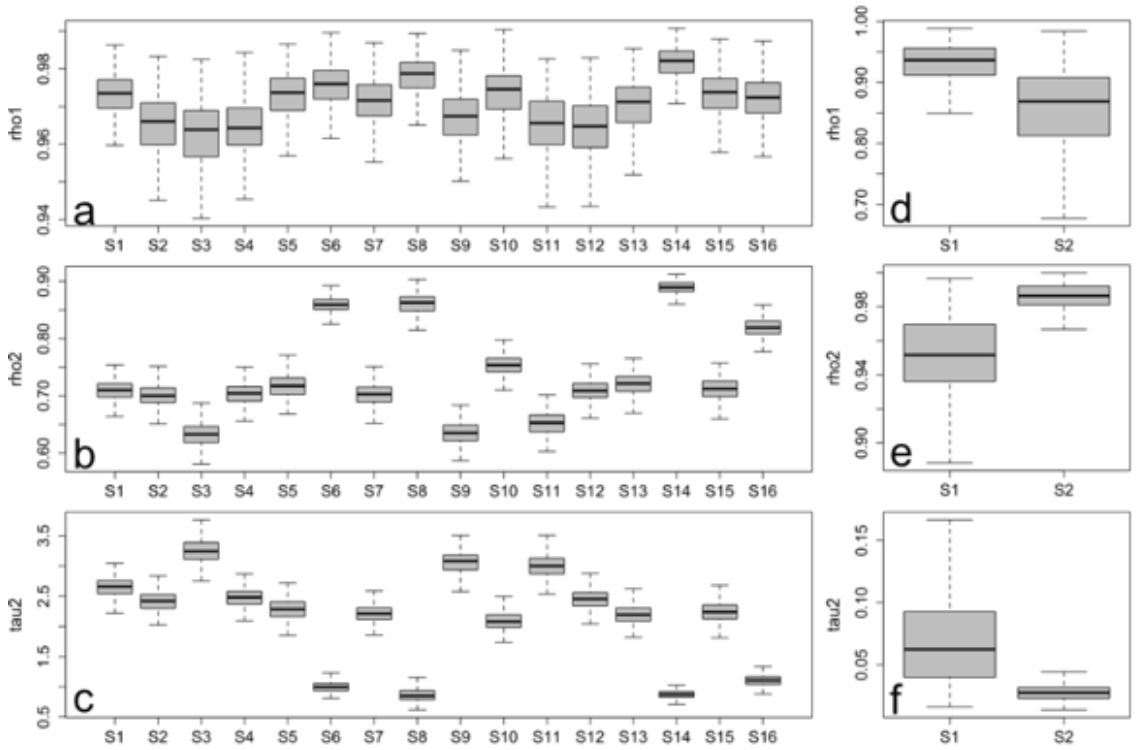

Figure 3.3: Box-plots of posterior samples of spatio-temporal random effects using STCARar. Each box plot corresponds to a scenario-specific model: (a) spatially autoregressive parameters, associated with simulation bias data-type; (b)-(c) the same as (a) but for temporally autoregressive parameters, and variances of spatial autocorrelations, respectively; (d)-(f) the same as (a)-(c), but associated with biofuel impact data-type.

3.5 Spatio-temporal modeling for scenario-combined data using STCARlinear

3.5.1 Simulation bias

By modeling the scenario-combined data using STCARlinear, the significance of seasons, observations, physics parameterizations, as well as their interactions was examined (Table 3.4). Each of the five candidate specifications (selected based on AIC, see Section 3.3.4, with parameters listed in Table 3.4) was evaluated using STCARlinear. We used a single chain with a total of 400,000 iterations. The number of iterations for the burnin period was 100,000, and the thinning rate was 100, so the number of samples used for the estimation of the parameters was 3000. As before, MCMC convergence was assessed using the Geweke diagnostic and the effective sample size was monitored as well. DIC values of the five specifications were 460476.7, 455997.2, 466952.4, 460424.7, and 464820.6, respectively. Therefore, Model2 offers the most satisfactory description of the data.

Table 3.4: Parameter lists of five candidate models, and the parameter estimates of candidates Model2 which have the smallest DIC for simulation bias data-type. Parameters included in each of the five candidate models are indicated with "Y", whereas not included with "N". In addition, significant parameters are colored in red.

	Parameter lists of candidate models					Summary of Model2				
Parameters	Model1	**Model2**	Model3	Model4	Model5	Median	0.025 CI	0.975 CI	effective sample size	Geweke diag
Intercept	Y	Y	Y	Y	Y	-2.23	-2.90	-1.56	52.70	-0.70
Spring-Winter	Y	Y	Y	Y	Y	0.42	0.39	0.46	3000.00	-0.20
Summer-Winter	Y	Y	Y	Y	Y	3.39	3.35	3.43	2733.30	-1.30
Fall-Winter	Y	Y	Y	Y	Y	2.70	2.66	2.74	3000.00	-2.10
spectral_nudging	Y	Y	Y	Y	Y	-0.45	-1.41	0.40	63.50	-0.30
cumulus	Y	Y	Y	Y	Y	-0.22	-1.05	0.95	51.10	0.60
microphysic	Y	Y	Y	Y	Y	1.56	0.80	2.15	64.00	1.70
(Spring-Winter)*spectral_nudging	N	Y	Y	Y	N	0.14	0.10	0.18	3000.00	0.60
(Summer-Winter)*spectral_nudging	Y	Y	Y	Y	Y	-0.41	-0.45	-0.37	3200.20	2.10
(Fall-Winter)*spectral_nudging	Y	Y	Y	Y	Y	-0.41	-0.45	-0.37	3298.50	1.40
(Spring-Winter)*cumulus	N	N	N	N	N					
(Summer-Winter)*cumulus	Y	Y	Y	Y	Y	-0.51	-0.54	-0.48	3000.00	0.80
(Fall-Winter)*cumulus	Y	Y	Y	Y	Y	-0.16	-0.19	-0.13	3000.00	0.40
(Spring-Winter)*microphysic	Y	Y	Y	N	Y	-0.83	-0.87	-0.79	3000.00	-1.30
(Summer-Winter)*microphysic	Y	Y	Y	Y	Y	-1.41	-1.45	-1.37	2496.80	-1.20
(Fall-Winter)*microphysic	Y	Y	Y	Y	Y	-1.07	-1.11	-1.03	3000.00	0.90
spectral_nudging*cumulus	Y	N	Y	Y	Y					
spectral_nudging*microphysic	Y	Y	N	Y	N	0.23	-1.03	1.49	66.80	0.20
cumulus*microphysic	N	N	N	N	N					
α	Y	Y	Y	Y	Y	-0.02	-0.04	0.01	3000.00	-0.70
τ_{int}	Y	Y	Y	Y	Y	2.37	2.25	2.50	3000.00	-0.90
τ_{slo}	Y	Y	Y	Y	Y	0.46	0.39	0.55	3000.00	0.90
ρ_{int}	Y	Y	Y	Y	Y	0.97	0.95	0.99	501.00	-0.20
ρ_{slo}	Y	Y	Y	Y	Y	0.95	0.92	0.98	3000.00	-0.50
DIC	460476.7	455997.2	466952.4	460424.7	464820.6					

However, some values of the Geweke diagnostic for Model2 are outside the acceptable range, such as factor (Fall-Winter) and interaction term (Summer-Winter)*spectral_nudging. In addition, the effective sample size for factors Intercept, spectral nudging, cumulus, microphysic, and spectral _nudging*microphysic are small. The above findings suggests our estimate could be unreliable. Given the length of MCMC and the associated computation burden, a more efficient estimation method is needed; this issue will be addressed in the next chapter.

Nevertheless, the range of posterior samples of spatio-temporal random effects using scenario-combined data are more concentrated than those of scenario-specific data (Figure 3.1 and Figure 3.4). Therefore, these posterior samples can be considered as weighted averages across all scenarios.

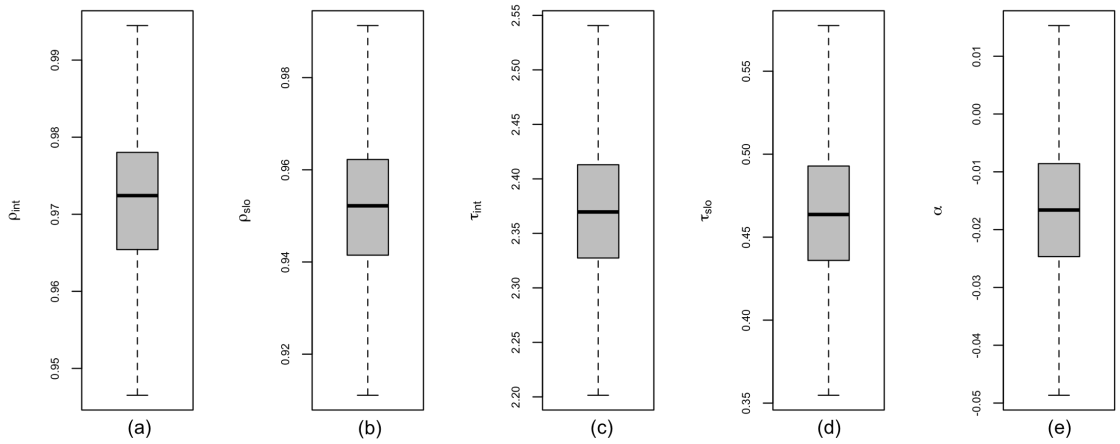

Figure 3.4: Box-plots of posterior samples of spatio-temporal random effects for simulation bias data-type using STCARlinear. (a) mean of spatially dependent intercept; (b)-(e) the same as (a) but for mean of spatially dependent slope, variance of spatially dependent intercept, variance of spatially dependent slope, and overall slope parameter, respectively.

3.5.2 Biofuel impacts

Five candidate models with different factors are selected to analyze the biofuel impact dataset using STCARlinear. A single chain with a total of 100,000 iterations is used for the MCMC. The number of iterations for the burnin period is 10,000, and the thinning rate is 100, so the number of samples used for the estimation of the parameters was 900. As before, the Geweke diagnostic is used to examine MCMC convergence; effective sample size was also monitored.

The seasonal factors are consistently included in each of the five candidate models, whereas the scenario factor appears only in Models 1, 3, and 5. With regard to interaction terms, the effect of (Summer-Winter)*scenario appears in all of the 5 candidate models, whereas the other interactions appear occasionally.

Among the five candidates, Model1 with fixed effects for seasons, a scenario indicator, and the (Summer-Winter)*scenario interaction term (i.e., x_1,x_2,x_3, x_4, and x_2x_4) resulted in the smallest DIC (84154, see Table 3.5). Seasonal factors and (Summer-Winter)*scenario interaction are statistically significant based on the 95% posterior credible interval. It can be observed that the averaged cooling impact associated with perennial bioenergy crops are 0.2°C and 0.3°C in Spring and Fall, respectively, relative to Winter. Comparing summer to winter, the simulated cooling impact is around 1.5°C using physics parameterizations scenario S1 (i.e., WDM6 for Microphysics, Grell 3D for cumulus scheme, and spectral nudging is applied); whereas 1.8°C cooling impact is estimated when simulated with Microphysics WSM3, Kain-Fritsch, and without spectral nudging. Overall, estimated temperature impacts associated with perennial bioenergy crops are robust across different physics parametizations, except for summer relative to winter. Similar to the simulation bias data, the range of posterior samples

of spatio-temporal random effects using scenario-combined data are more concentrated than the ones derived from scenario-specific data (Figure 3.1 and Figure 3.5). The above results can be considered reliable, as the values of Geweke diagnostics are within the acceptable range and the effective sample size is satisfactory for all parameters (Table 3.5).

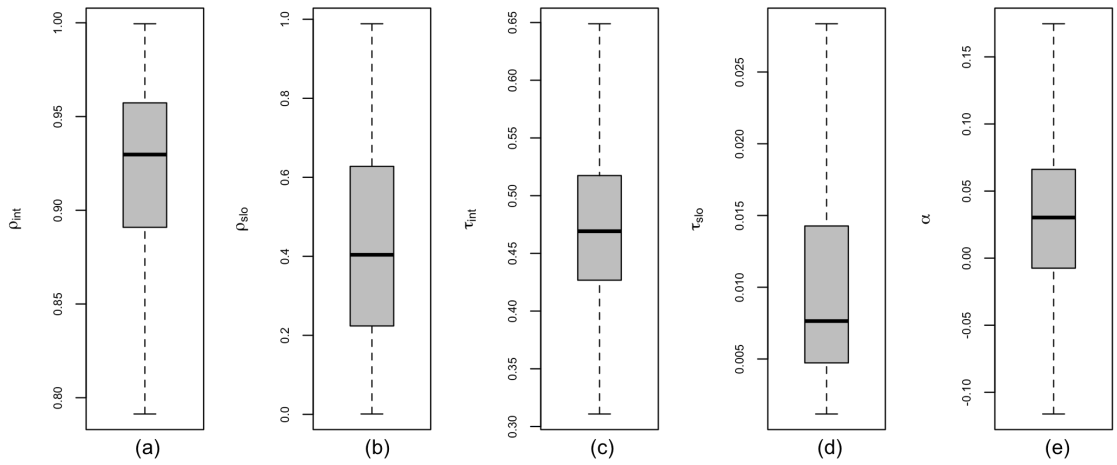

Figure 3.5: Box-plots of posterior samples of spatio-temporal random effects for biofuel impact dataset using STCARlinear. (a) mean of spatially dependent intercept; (b)-(e) the same as (a) but for mean of spatially dependent slope, variance of spatially dependent intercept, variance of spatially dependent slope, and overall slope parameter, respectively.

Table 3.5: Parameter lists of five candidate models, and the parameter estimates of candidates Model1 which have the smallest DIC for biofuel data-type. Parameters included in each of the five candidate models were indicated with "Y", whereas not included with "N". In addition, significant parameters were colored in red.

	Parameter lists of candidate models					Summary of Model2				
Parameters	**Model1**	Model2	Model3	Model4	Model5	Median	0.025 CI	0.975 CI	effective sample size	Geweke diag
Intercept	Y	Y	Y	Y	Y	-0.1106	-0.3744	0.2102	259.2	1.5
Spring-Winter	Y	Y	Y	Y	Y	-0.1894	-0.2733	-0.0964	900	-0.2
Summer-Winter	Y	Y	Y	Y	Y	-1.8492	-1.9608	-1.7295	915.4	0.4
Fall-Winter	Y	Y	Y	Y	Y	-0.3185	-0.404	-0.2313	1087	0.1
scenario_dummy	Y	N	Y	N	Y	0.0697	-0.5625	0.6102	256.4	-1.5
(Spring-Winter)*scenario_dummy	N	N	N	Y	Y					
(Summer-Winter)*scenario_dummy	Y	Y	Y	Y	Y	0.3298	0.1967	0.4747	900	-1
(Fall-Winter)*scenario_dummy	N	Y	Y	Y	N					
alpha	N	N	N	N	N	0.0303	-0.0764	0.1298	900	0.9
tau2.int	N	N	N	N	N	0.4692	0.3639	0.6098	900	-0.7
tau2.slo	N	N	N	N	N	0.0076	0.0022	0.063	158	2
nu2	N	N	N	N	N	4.5651	4.4721	4.6575	900	1.1
rho.int	N	N	N	N	N	0.9297	0.7954	0.9885	900	0.2
rho.slo	N	N	N	N	N	0.4038	0.0434	0.9032	264	-0.3
DIC	84165	84302.8	84321.6	84308.6	84283.7					

3.6 Concluding remarks

In this chapter, WRF simulated temperatures associated with control simulation bias, as well as biofuel impacts, were modeled using spatio-temporal bayesian hierarchical models. Our findings suggest that models with spatially varying intercepts and slopes can offer a satisfactory description of the spatio-temporal dependence of the data. The simulated cooling impact associated with perennial bioenergy crops differ by seasons significantly. Most importantly, simulated impacts on temperatures due to perennial bioenergy crop expansion are found robust to physics parameterizations. This robustness, however, does not hold in summer relative to winter.

This work has several limitations. One of them is that sensitivity analysis of prior distributions is not performed: different prior specifications may result in different inferences. Besides, parameter estimation techniques (i.e., MCMC vs INLA) are not compared in depth. A more thorough comparison of estimation accuracy and computation times should be considered. In addition, issues related to change of support and alignment were ignored at the pre-processing data stage. Most importantly, MCMC simulation requires a large number of iterations to achieve convergence, for the BHM examined. Therefore, a more efficient modeling approach for large spatio-temporal datasets should be considered.

It is worth noting that the physics parameterizations and observations are included in the models as fixed effects under the assumption of spatial and temporal homogeneity. However, it is possible that spatially varying effects exist (Kang et al. 2012). Moreover, multivariate hierarchical spatio-temporal modeling (i.e., for temperature and precipitation simultaneously) should be performed as the aforementioned variables are both significant for climate model comparison. Despite the above-mentioned

limitations, this work establishes a framework to quantitatively assess the impact of physics parameterizations on WRF simulation, focusing on an application associated with perennial bioenergy crops expansion.

Chapter 4

SPATIO-TEMPORAL MODELING FOR REGIONAL CLIMATE MODEL EVALUATION: EIGENVECTOR FILTERING VERSUS BAYESIAN CAR

4.1 Introduction

A suite of 10-year ensemble-based simulations was conducted to investigate the hydroclimatic impacts due to large-scale deployment of perennial bioenergy crops across the continental United States (Wang *et al.* 2017). Given the deterministic nature of the simulations, uncertainties of hydroclimatic impacts caused by physics parameterizations exist within the ensemble. To better examine the robustness of impacts on climate and hydrology associated with bioenergy crops expansion, Bayesian hierarchical spatio-temporal statistical modelling (BHM) has been implemented (see results in Chapter 3). However, BHM were estimated based on MCMC, which may take a long time to converge, especially for large datasets.

The simulated data are lattice data that are correlated in space and time. To take into account spatial correlation, one popular approach is eigenvector spatial filtering (ESF; Griffith 1996; 2000; 2003), which is mathematically associated with Moran coefficients (Moran 1948; Tiefelsdorf and Griffith 2007; Griffith and Paelinck 2011; Chun and Griffith 2013; Cressie and Wikle 2015). ESF models spatial dependencies via including proxy variables in the standard linear (Griffith 2003) or generalized linear regression framework (Griffith 2002; 2004b; Chun 2008). These proxies are a subset of the orthogonal and uncorrelated eigenvectors which are constructed based on the available (through a weights matrix) spatial neighborhood information (Pace *et al.*

2013; Griffith 2004a; Griffith and Peres-Neto 2006; Thayn and Simanis 2013; Griffith and Fischer 2016; Griffith 2011; Thayn and Simanis 2013). Statistical properties of unbiasedness, efficiency, and consistency are held by the ESF estimators (Chun and Griffith 2014; Griffith 2017). At least two alternative specifications of proxy variables can be constructed, leading them to be either correlated or orthogonal with the explanatory variables. Using the latter type of proxy variables, Hughes and Haran (2013) extended ESF with a random effects specification (RE-ESF), which took into account spatial confounding, i.e., the proposed method mitigated the variance inflation due to the collinearity between explanatory variables and a latent spatial process (Reich *et al.* 2006; Hodges and Reich 2010). Murakami and Griffith (2015) further improved RE-ESF by introducing a computationally efficient REML estimation scheme and by examining the effects of scale in the spatial dependency structure. Hefley *et al.* (2017) optimize predictive ability of the RE-ESF and jointly regularize the regression coefficients and spatial random effects. Recently, spatially varying coefficients within the ESF framework are also studied (Helbich and Griffith 2016; Murakami *et al.* 2017).

When temporal information should also be considered, ESF coupled with generalized linear mixed models (GLMM) have been studied to incorporate both spatial and temporal components efficiently and effectively (Chun and Griffith 2011; Patuelli *et al.* 2011; Chun 2014). In this specification, spatial autocorrelation is captured by adding orthogonal and uncorrelated eigenvectors as specified in ESF, while temporal components are captured as random effects in GLMM. Chun (2014) also studied ESF associated with GLMMs by simultaneously allowing spatial and temporal correlation structures. Recently, conventional ESF has been extended to eigenvector space-time filtering via specifying a matrix that summarizes spatio-temporal neighborhood relationships (i.e., spatio-temporal contemporaneous structure, Griffith and Chun 2015).

However, the discussion about this model is very limited; in addition, the efficiency and performance of this approach has not been examined.

The efficiency and performance of ESF can also be influenced by the way of selecting the most relevant candidates from a large number of eigenvectors. These selected eigenvectors capture different scales of spatial autocorrelation in the residuals, leading the remaining part of the residuals to be uncorrelated. Some strategies of eigenvector selection have been proposed and studied. Tiefelsdorf and Griffith (2007) proposed a semiparametric approach based on the criterion of MC/MCmax > 0.25, where MCmax is the largest positive MC value. However, the choice of this criterion appears to be subjective. Stepwise regression (Griffith 2003) has also been applied to select the most significant eigenvectors. Nevertheless, stepwise selection results may be inconsistent, affected by stochastic errors, and may not reach global optimal solution (Fan and Li 2001; Shen and Ye 2002; Whittingham *et al.* 2006).

Alternatively, Seya *et al.* (2015) investigated the least absolute shrinkage and selection operator (Lasso), a penalized estimator, as a faster and more reliable eigenvector selection procedure. However, problems related to Lasso are that it could shrink and select eigenvectors according to the same tuning weight, and that it does not possess oracle properties (i.e., being consistent in parameter estimation and variable selection) in some circumstances when the predictors are significantly correlated (Zou 2006).

In this chapter, we propose a framework of spatio-temporal eigenvector filtering (hereafter: STEF). Spatial and temporal effects are modeled simultaneously based on a spatio-temporal contemporaneous structure. Three approaches of introducing proxy variables to the model are specified in terms of dealing with the spatial confounding problem: proxy variables with spatial confounding, proxy variables without spatial confounding, and proxy variables with intermediate spatial confounding (hereafter,

STEF_CF, STEF_NCF, and STEF_ICF, respectively). For each approach, two alternative two-stage modeling strategies are used. In addition, Adaptive LASSO, a regularization method which avoids overfitting and possesses oracle properties, is used to select significant eigenvectors. Specifically, a fast approximation of Adaptive LASSO estimates – the Least Squares Approximation (LSA) to Adaptive LASSO is implemented in the model to decrease computational time. For STEF_ICF, Variance inflation factor (VIF) based filtering and Sure Independence Screening (SIS) algorithms are applied before LSA to alleviate spatial confounding and improve LASSO performance. Monte Carlo experiments are conducted using the proposed methods. In addition, STEFs are applied to quantify the robustness of simulated hydroclimatic impacts associated with bioenergy crops to alternative physics parameterizations.

This chapter is arranged as follows. Section 4.2 presents a review of the ESF approach and its extensions for spatio-temporal correlation structures; in addition, adaptive LASSO is discussed. The STEF approaches are introduced in Section 4.3 and Monte Carlo simulation experiments are presented in Section 4.4. The application to the robustness of hydroclimatic impacts associated with perennial bioneregy crops expansion is studied in Section 4.5. Concluding remarks and suggestions for future work are presented in Section 4.6.

4.2 Methodology

4.2.1 The eigenvector spatial filtering (ESF) approach

4.2.1.1 The Moran coefficient (MC)

MC is a diagnostic statistic for spatial dependence, which is formulated as (see, Anselin and Rey 1991):

$$\mathrm{MC}[\boldsymbol{y}] = \frac{N}{\mathbf{1}'\boldsymbol{C}\mathbf{1}} \frac{\boldsymbol{y}'\boldsymbol{M}\boldsymbol{C}\boldsymbol{M}\boldsymbol{y}}{\boldsymbol{y}'\boldsymbol{M}\boldsymbol{y}}. \tag{4.1}$$

N represents sample size, $\mathbf{1}$ is a $N \times 1$ vector of ones, $\boldsymbol{C}$ is a symmetric connectivity matrix whose diagonal elements are zero, $\boldsymbol{M}$ is a projection matrix, and $\boldsymbol{y}$ is a vector of values of georeferenced data. Two alternative types of projection matrix $\boldsymbol{M}$, $\boldsymbol{M}_{(1)} = \boldsymbol{I} - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}' = \boldsymbol{I} - \dfrac{\mathbf{1}\mathbf{1}'}{N}$, and $\boldsymbol{M}_{(X)} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' = \boldsymbol{I} - \boldsymbol{P}_X$, can be specified, where $\boldsymbol{I}$ is an $N \times N$ identity matrix, and $\boldsymbol{X}$ is a $N \times K$ matrix of K explanatory variables.

The expectation of MC is:

$$E\left[\mathrm{MC}\right] = \begin{cases} -\dfrac{1}{N-1} & \text{for } \boldsymbol{M}_{(1)}, \\ -\dfrac{N}{\mathbf{1}'\boldsymbol{C}\mathbf{1}} \dfrac{tr\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{C}\boldsymbol{X}\right]}{N-K-1} & \text{for } \boldsymbol{M}_{(X)}. \end{cases} \tag{4.2}$$

where tr is the trace of a matrix. MC $>$ E[MC], MC $<$ E[MC], and MC $=$ E[MC] imply positive, negative, and no spatial dependence, respectively; hence, MC is positive if the values in $\boldsymbol{y}$ display positive spatial dependence and negative if they demonstrate negative spatial dependence.

The ESF approach accounts for global and local spatial autocorrelation in the residuals; subsequently, ordinary least squares estimates can be computed as in

an i.i.d. setting. Moran ESF is based on the MC. Let $\{\boldsymbol{e}_1, \dots \boldsymbol{e}_N\}_{(1)}$ be a set of eigenvectors of $\boldsymbol{M}_{(1)}\boldsymbol{C}\boldsymbol{M}_{(1)}$ in Eq. 4.1, and $\{\lambda_1, \dots, \lambda_N\}_1$ be the set of corresponding eigenvalues. Similarly, let $\{\boldsymbol{e}_1, \dots \boldsymbol{e}_N\}_{(X)}$ be a set of eigenvectors of $\boldsymbol{M}_{(X)}\boldsymbol{C}\boldsymbol{M}_{(X)}$, and $\{\lambda_1, \dots, \lambda_N\}_X$ be the set of corresponding eigenvalues. ESF utilizes eigenvectors $\{\boldsymbol{e}_1, \dots \boldsymbol{e}_N\}_{(1)}$ or $\{\boldsymbol{e}_1, \dots \boldsymbol{e}_N\}_{(X)}$, which are mutually uncorrelated and orthogonal. Each eigenvector is associated with a certain degree of latent spatial dependence, representing global to local map patterns (Tiefelsdorf and Griffith 2007). Let $\boldsymbol{E}_{(1)}$ and $\boldsymbol{E}_{(X)}$ be $N \times N$ matrices which are composed of elements of $\{\boldsymbol{e}_1, \dots \boldsymbol{e}_N\}_{(1)}$ and $\{\boldsymbol{e}_1, \dots \boldsymbol{e}_N\}_{(X)}$, respectively. $\boldsymbol{E}_{(X)}$ contains columns which are orthogonal to the explanatory variables in X, whereas the columns in $\boldsymbol{E}_{(1)}$ are potentially correlated with the columns in X (Griffith 2003).

To interpret the eigenvectors in terms of the MC coefficient, $\boldsymbol{MCM}$ can be eigen-decomposed as $\boldsymbol{MCM} = \boldsymbol{E}\boldsymbol{\Lambda}\boldsymbol{E}'$. In this case $\boldsymbol{E}$ is used to represent $\boldsymbol{E}_{(1)}$ or $\boldsymbol{E}_{(X)}$ (i.e., the i^{th} column of $\boldsymbol{E}$ corresponds to the i^{th} eigenvector $\boldsymbol{e}_i$). In addition, $\boldsymbol{\Lambda}$ represents a $N \times N$ diagonal matrix, with the i^{th} element on the diagonal being the i^{th} eigenvalue λ_i. Thus the MC of $\boldsymbol{e}_i$ is represented as (Murakami and Griffith 2017):

$$\begin{aligned} \mathrm{MC}[\boldsymbol{e}_i] &= \frac{N}{\mathbf{1}'\boldsymbol{C}\mathbf{1}} \frac{\boldsymbol{e}_i'\boldsymbol{MCM}\boldsymbol{e}_i}{\boldsymbol{e}_i'\boldsymbol{M}\boldsymbol{e}_i} \\ &= \frac{N}{\mathbf{1}'\boldsymbol{C}\mathbf{1}} \frac{\boldsymbol{e}_i'\boldsymbol{E}\boldsymbol{\Lambda}\boldsymbol{E}'\boldsymbol{e}_i}{\boldsymbol{e}_i'\boldsymbol{e}_i} \\ &= \frac{N}{\mathbf{1}'\boldsymbol{C}\mathbf{1}} \lambda_i \end{aligned} \tag{4.3}$$

Therefore, the eigenvectors can be interpreted as follows: "The first eigenvector, $\boldsymbol{e}_1$, is the set of real numbers that has the largest MC value achievable by any set for the geographical arrangement defined by the spatial connectivity matrix $\boldsymbol{C}$; the second eigenvector is the set of real numbers that has the largest achievable MC by any set that is orthogonal (hence uncorrelated) with $\boldsymbol{e}_1$; the third eigenvector is the third such set of real numbers; and so on, through $\boldsymbol{e}_N$, the set of real numbers that has

the largest negative MC achievable by any set that is orthogonal with the preceding $N-1$ eigenvectors" (Griffith 2003).

4.2.1.2 Conventional/standard ESF

The ESF regression model is formulated as:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{E}\boldsymbol{\gamma} + \boldsymbol{\epsilon},\ \epsilon \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}) \tag{4.4}$$

where $\boldsymbol{y}$ is a $N \times \boldsymbol{1}$ vector of response variable values; $\boldsymbol{X}$ is a $N \times K$ matrix of explanatory variables; $\boldsymbol{E}$ is a $N \times L$ matrix composed of a subset of L eigenvectors (L $<$ N) from $\boldsymbol{M}_{(1)}\boldsymbol{C}\boldsymbol{M}_{(1)}$ or $\boldsymbol{M}_{(X)}\boldsymbol{C}\boldsymbol{M}_{(X)}$; $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are parameter vectors whose sizes are $K \times 1$ and $L \times 1$, respectively; $\boldsymbol{\epsilon}$ represents a Normally distributed error term, with variance σ^2.

4.2.2 Space-time eigenvector filter (STEF) framework

Extending the idea of ESF from spatial to spatio-temporal phenomena, STEF was introduced (Griffith 2012; Griffith and Chun 2015; Griffith and Paelinck 2018) based on the latent structure of the space-time Moran Coefficient (Cliff and Ord 1981; Griffith 1981). Instead of using a spatial weight matrix $\boldsymbol{C}$, in this case a contemporaneous spatio-temporal dependence matrix was considered:

$$\boldsymbol{C}_{ST} = \boldsymbol{I}_T \otimes \boldsymbol{C}_S + \boldsymbol{C}_T \otimes \boldsymbol{I}_S \tag{4.5}$$

$\boldsymbol{I}_T$ is a $T \times T$ identity matrix, $\boldsymbol{I}_S$ is a $N \times N$ identity matrix, $\boldsymbol{C}_S$ is the $N \times N$ spatial weight matrix, $\boldsymbol{C}_T$ is a $T \times T$ temporal weight matrix, with upper and lower off-diagonal elements equal to 1 and zeroes elsewhere. Therefore $\boldsymbol{C}_{ST}$ is of dimension

$n \times n$ where $n = N \times T$. Using this specification, the spatio-temporal contemporaneous structure assumes that a value at a given location for a particular point in time, is associated with a value at that location for the previous point in time and the values of nearby locations for the same point in time (Fig. 4.1).

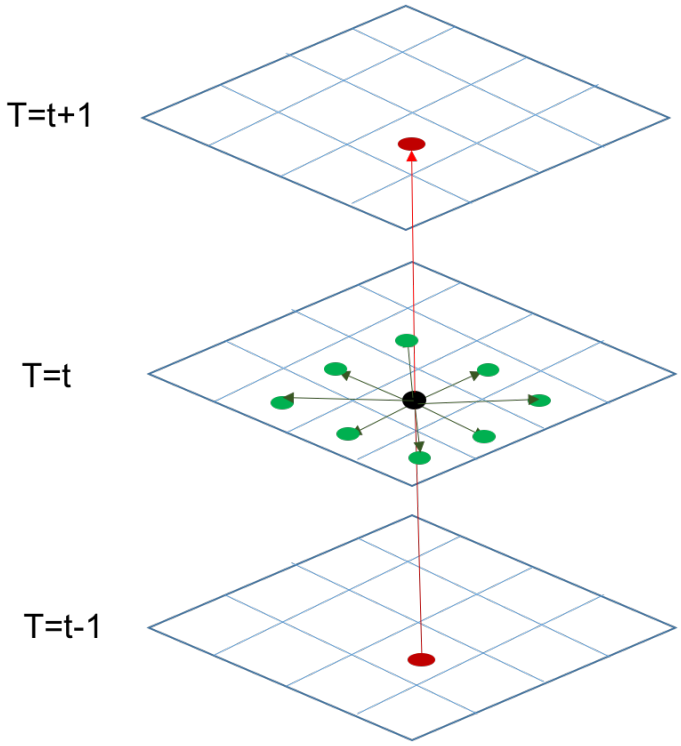

Figure 4.1: Spatio-temporal contemporaneous specification. Black dot represents the value at a specific location at time t; green dots represent the associated instantaneous values at neighboring locations; red dots represent the associated values at the same location at time $t-1$ and $t+1$.

$\boldsymbol{C}_{ST}$ in Eq. 4.5 can be considered as a more general form of weight matrix (Griffith and Paelinck 2018). If $T = 1$,

$$\boldsymbol{C}_{ST} = 1 \otimes \boldsymbol{C}_S + 0 \otimes \boldsymbol{I}_S = \boldsymbol{C}_S \tag{4.6}$$

In this case, $\boldsymbol{C}_{ST}$ corresponds to the conventional MC, which is static.

In this chapter, the spatial weight matrix $\boldsymbol{C}_S$ in Eq. 4.5 is defined using the queen's adjacency criterion. Instead, rook's adjacency criterion is used for STEF in other studies of (i.e., Griffith 2012; Griffith and Chun 2015; Griffith and Paelinck 2018). To further examine the influence of weight matrices on the results of STEF, a Monte Carlo experiment is conducted in section 4.4.

Analogous to the preceding discussions of ESF, STEF adds a set of synthetic proxy variables as control variables into a regression model. These proxies are selected eigenvectors associated with a space-time contemporaneous connectivity matrix, which connects elements together in both space and time. Hence these eigenvectors can be interpreted as follows:

"The first eigenvector, say $\boldsymbol{e}_1$, is the set of real numbers that has the largest space-time MC achievable by any set for the areal unit articulation defined by the space-time connectivity matrix $\boldsymbol{C}_{ST}$; the second eigenvector is the set of real numbers that has the largest achievable space-time MC by any set that is orthogonal (hence uncorrelated) with $\boldsymbol{e}_1$; the third eigenvector is the third such set of real numbers; and so on through $\boldsymbol{e}_L$, the set of real numbers that has the largest negative space-time MC achievable by any set that is orthogonal and uncorrelated with the preceding $N-1$ eigenvectors" (Griffith 2012).

In general, the STEF regression model is formulated as:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{E}\boldsymbol{\gamma} + \epsilon,\ \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}) \tag{4.7}$$

where $\boldsymbol{y}$ is a $n \times \boldsymbol{1}$ vector of response values, with $n = N \times T$, given $\boldsymbol{y}$ with N areal units and T temporal units; $\boldsymbol{X}$ is a $n \times K$ matrix of explanatory variables; $\boldsymbol{E}$ is a $n \times L$ matrix composed of a subset of L eigenvectors (L < n, the choice of L is discussed in the next section) based on $\boldsymbol{M}\boldsymbol{C}_{ST}\boldsymbol{M}$; $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are parameter vectors

whose sizes are $K \times 1$ and $L \times 1$, respectively; $\boldsymbol{\epsilon}$ represents a Normally distributed error term, with variance σ^2.

4.2.2.1 Eigenvector selection

ESF often constructs an excessive number of eigenvectors: several strategies have been proposed aiming to reduce the number of eigenvectors to obtain a more parsimonious model. Particularly, Seya *et al.* (2015) studied the use of the least absolute shrinkage and selection operator (Lasso) to select significant eigenvectors. The Lasso is a regularization technique for simultaneous estimation and variable selection (Tibshirani 1996). In this case coefficient estimates are derived by solving the following optimization problem:

$$\hat{\boldsymbol{\beta}}_{lasso} = argmin_{\beta} \left(\sum_{i=1}^{N} (y_i - \boldsymbol{x_i'}\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} \mid \beta_j \mid \right) \tag{4.8}$$

with λ representing a nonnegative regularization parameter. The second term in Eq. 4.8 is the so-called "penalty", which is crucial for the success of the technique: the Lasso continuously shrinks the coefficients towards 0 as λ increases, and some coefficients are shrunk to 0 exactly if λ is sufficiently large. This procedure takes into account the bias-variance trade-off, leading to high predictive accuracy relative to conventional least squares estimates. However, depending on the correlation structure of the predictors, Lasso may not possess the oracle properties (Meinshausen and Bühlmann 2006).

Adaptive Lasso, an alternative penalized estimator which is similar to Lasso, is based on coefficient-specific penalties. Adaptive Lasso possesses the same advantage as the Lasso: it performs parameter estimation and feature selection via continuously shrinking some coefficients to zero. Furthermore, it possesses the oracle properties

given a suitable tuning parameter λ and a consistent initial estimator. In a linear regression setting, the Adaptive Lasso seeks to minimize (Zou 2006; Wang and Leng 2007):

$$\hat{\beta}_{alasso} = argmin_{\beta} \left(\sum_{i=1}^{N} (y_i - \boldsymbol{x_i'}\boldsymbol{\beta})^2 + \sum_{j=1}^{p} \lambda_j \mid \beta_j \mid \right) \tag{4.9}$$

In this chapter, Least squares approximation(LSA) to Adaptive Lasso is implemented, since it provides efficient computation of the Adaptive Lasso estimates (Wang and Leng 2007). This approach uses Least Angle Regression (LARS) algorithm to find the solution path of Adaptive Lasso, at the computational cost of a single OLS fit (Efron *et al.* 2004). In addition, LSA estimator is as efficient as the oracle asymptotically, as long as the tuning parameters are selected appropriately (Wang and Leng 2007).

4.2.2.2 Spatial confounding alleviation

The previous strategies can be applied to select significant eigenvectors from $\{\boldsymbol{e}_1, \ldots \boldsymbol{e}_N\}_{(X)}$, as these eigenvectors are mutually uncorrelated and orthogonal to the explanatory variables $\boldsymbol{X}$. However, when the set of eigenvectors $\{\boldsymbol{e}_1, \ldots \boldsymbol{e}_N\}_{(1)}$ based on $\boldsymbol{M}_{(1)}\boldsymbol{C}\boldsymbol{M}_{(1)}$ is used, some columns in $\boldsymbol{X}$ might be strongly correlated with these eigenvectors (i.e., causing multicolinearity), leading to variance inflation in parameter estimation. Although $\boldsymbol{M}_{(1)}\boldsymbol{C}\boldsymbol{M}_{(1)}$ does not take into account spatial confounding, according to Griffith (2017) the confidence intervals estimated using $\boldsymbol{M}_{(1)}\boldsymbol{C}\boldsymbol{M}_{(1)}$ may be more accurate.

Filtering analyses based on variance inflation factors (VIF) are applied to alleviate severe multicollinearity. These filtering analyses consist of two components: filtering for explanatory variables and filtering for eigenvectors. First of all, stepwise selection

of explanatory variables based on VIF is used to remove collinear predictors. This procedure computes VIF for each explanatory variable using the whole set of predictors; then, the predictor with the highest VIF is removed if a VIF-threshold is exceeded. VIF is calculated again using the reduced set of predictors and the predictor with the highest VIF is removed if a VIF-threshold is exceeded. The previous step is repeated until all predictors possess VIFs which are below a pre-specified threshold [10 in this Chapter, which corresponds to $R^2 = 0.9$ in Eq.4.10 below]. Specifically, VIF is calculated as:

$$\text{VIF} = \frac{1}{1 - R^2} \tag{4.10}$$

where R^2 is the coefficient of determination of the predictor-specific regression model which includes all the remaining predictors in the explanatory part.

For filtering analysis on eigenvectors, each eigenvector is regressed on all the explanatory variables. Since eigenvectors are mutually orthogonal, the rest of eigenvectors are not included in the regression for VIF calculation. Then all the VIF values for each specific eigenvector (Eq. 4.10) are calculated, where R^2 is the coefficient of determination of the eigenvector-specific regression equation. If the value of VIF is greater than 10, the corresponding eigenvector will be considered highly collinear with the fixed effects and it will be removed.

In addition, sure independence screening (vanilla SIS, or SIS) is implemented after VIF filtering to further eliminate non-significant eigenvectors (Fan and Lv 2008; Barut *et al.* 2016). SIS eliminates predictors with low marginal correlation with the response, reducing the dimension of the dataset (Fan and Lv 2008), thus improving the efficiency of the Adaptive Lasso estimator. In general, SIS first ranks the VIF-selected p eigenvectors based on their marginal correlations $\widehat{corr}((\boldsymbol{e}_{(i)}, \boldsymbol{y})$, where $i \in \{1, \ldots, p\}$; it then retains d eigenvectors which have the corresponding d largest correlation. Fan

and Lv (2008) proposed the following threshold:

$$d = n/\log\, n, \tag{4.11}$$

where $n = N \times T$ is the sample size of the data. However, since there may still exists correlation between fixed effects and selected eigenvectors and data may still be autocorrelated in space and time, effective sample size $n^\star$ should be used instead of conventional sample size n. According to Griffith (2005),

$$n^\star = (1 - R^2)n. \tag{4.12}$$

In this case, R^2 is the goodness of fit metric for the regression model that contains the residuals of $\boldsymbol{y}$ with VIF-filtered predictors as its response and all the significant eigenvectors in the explanatory part. Finally, $d^\star$ eigenvectors will be selected based on

$$d^\star = n^\star/\log\, n^\star. \tag{4.13}$$

Using VIF-SIS, significant eigenvectors will be selected appropriately, when spatial confounding exists. The VIF-SIS filtering procedure is summarized in Algorithm 1.

4.2.2.3 STEF specifications

Based on the previous discussion, three approaches of introducing proxy variables can be specified in order to address the spatial confounding problem: proxy variables which allow spatial confounding, proxy variables which do not allow spatial confounding, and proxy variables which partially allow spatial confounding, but eliminate strong collinearity between Moran eigenvectors and predictors (i.e., STEF_CF, STEF_NCF, and STEF_ICF, respectively). More specifically, in what follows STEF_CF is specified using only the positive eigenvectors of $\boldsymbol{M}_{(1)}\boldsymbol{C}_{ST}\boldsymbol{M}_{(1)}$ (The suggestion of using

Algorithm 1 VIF-SIS filtering

1: Inputs: All eigenvectors $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n\}$ from $\boldsymbol{M}_{(1)}\boldsymbol{C}_{ST}\boldsymbol{M}_{(1)}$.
2: Perform VIF filtering from (4.10) on X (i.e., K explanatory variables) using stepwise selection, obtain a modified (filtered) X including r variables that do not have severe multicollinearity, where $r \in \{1, \ldots, K\}$.
3: For every $i \in \{1, \ldots, n\}$, perform VIF filtering from (4.10) for $\boldsymbol{e}_i$ using the modified (filtered) X based on Step 2, and keep only p $\boldsymbol{e}_i s$ that do not have severe multicollinearity, where $p \in \{1, \ldots, n\}$.
4: For every $j \in \{1, \ldots, p\}$, calculate marginal correlation of eigenvector $\widehat{corr}((\boldsymbol{e}_{(i)}, \boldsymbol{y})$.
5: Rank all the p marginal correlations and retain the top d eigenvectors from (4.11) with the largest absolute correlations.
6: Obtain multiple correlation R^2 for $\boldsymbol{y}$ regressed on the selected d eigenvectors.
7: Obtain effective sample size $n^\star$ from (4.12).
8: Obtain $d^\star$, the modified number of eigenvectors being selected, from (4.13).
9: Outputs: $d^\star$ eigenvectors $\{\boldsymbol{e}_{(1)}, \ldots, \boldsymbol{e}_{(d^\star)}\}$ are selected.

only positive eigenvectors can be found in Hughes and Haran (2013)); STEF_NCF is also based on positive eigenvectors, using $\boldsymbol{M}_{(X)}\boldsymbol{C}_{ST}\boldsymbol{M}_{(X)}$; STEF_ICF, is based on all eigenvectors of $\boldsymbol{M}_{(1)}\boldsymbol{C}_{ST}\boldsymbol{M}_{(1)}$ but VIF-SIS (see Algorithm 1) is implemented first to remove collinear explanatory variables, problematic eigenvectors which contribute to severe confounding, and irrelevant eigenvectors which reduce the efficiency of Lasso-type estimators.

For each approach of introducing proxy variables, two alternative two-stage STEF modeling strategies are used. In the first method (named method1), the LSA approximation to adaptive Lasso is applied first; both explanatory variables and eigenvectors are selected in this step. However, given that we are interested in estimating the coefficients of the explanatory variables, we only considered the selection of the eigenvector part, while keeping the explanatory variables in our model. Explanatory variables and selected eigenvectors are then used in a conventional linear regression framework to make statistical inference. This least squares after Lasso type of model estimation

has been shown to possess equivalent rates of convergence compared to Lasso, but smaller bias (Belloni *et al.* 2013).

On the contrary, the second method (i.e., method2) first obtains residuals from the covariates-only model and then, using these residuals as input, STEF is fitted to select eigenvectors that best capture the spatio-temporal autocorrelation structure, based on the LSA approximation to adaptive Lasso. Combining these eigenvectors with covariates, the regression model which using $\boldsymbol{y}$ as input is fitted to obtain estimates of fixed effects coefficients.

Combining different methods for creating proxy variables with alternative modeling strategies, we end up with six alternative STEF spefications; namely, STEF_CF_method1, STEF_CF_method2, STEF_NCF_method1, STEF_NCF_method2, STEF_ICF_method1, and STEF_ICF_method2. In this chapter, the performance of STEF ensemble modeling will be examined in a series of Monte Carlo experiments.

4.3 Monte Carlo simulations

4.3.1 Simulated data

In this section, Monte Carlo experiments are performed in order to examine the capability of capturing the true coefficients and the true spatio-temporal structure of the data. In all Monte Carlo experiments, the spatial domain is a 10×10 lattice, with coordinates of the domain vertices restricted to the unit square, and $t = 1, \ldots, 20$ time periods. Thus there are 100 areal units for 20 consecutive time period and the total sample size is 2000. These units possess spatio-temporal contemporaneous structure

(Eq. 4.5), which is formed by a combination of a spatial weight matrix and a temporal weight matrix (Fig. 4.2).

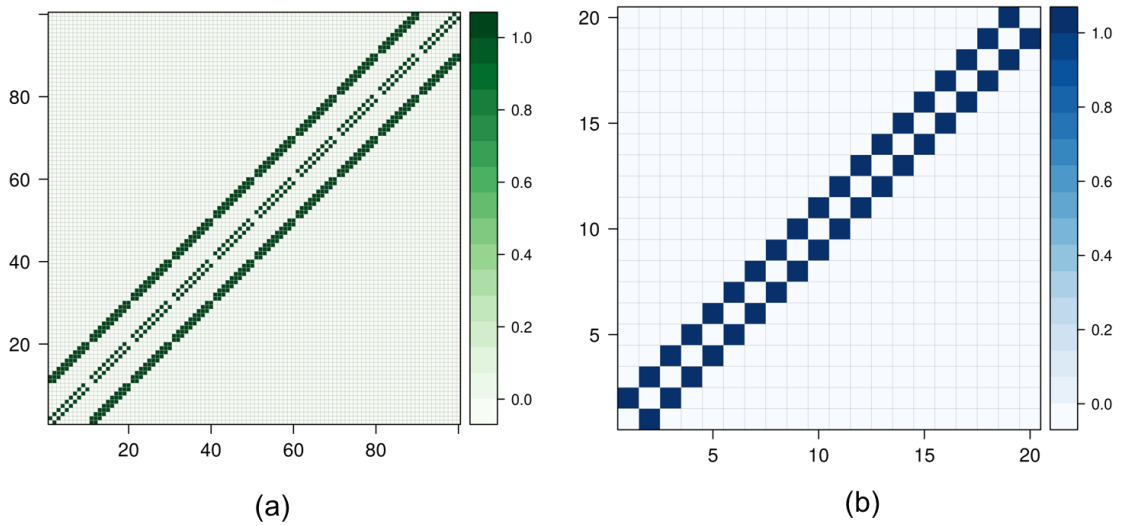

Figure 4.2: Visualization of the (a) spatial weight matrix for a 10×10 lattice, and (b) temporal weight matrix for 20 consecutive time period.

The design matrix for the explanatory variables are chosen to be $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2]$, with values of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ given by the longitude and latitude coordinates of the lattice within the unit square, respectively. Each explanatory variable (i.e., $\boldsymbol{x}_1$ or $\boldsymbol{x}_2$) is a vector of length 2000×1, ordered by the first 100 data points to be the set of all 100 spatial locations at time 1, the next 100 are the set of spatial points for time 2 and so on. Specifically, $\boldsymbol{x}_1 = \left[x_{1_{1,1}}, \ldots, x_{1_{100,1}}, x_{1_{1,2}}, \ldots, x_{1_{100,2}}, \ldots, x_{1_{1,20}}, \ldots, x_{1_{100,20}}\right]'$ and $\boldsymbol{x}_2 = \left[x_{2_{1,1}}, \ldots, x_{2_{100,1}}, x_{2_{1,2}}, \ldots, x_{2_{100,2}}, \ldots, x_{2_{1,20}}, \ldots, x_{2_{100,20}}\right]'$. Since coordinates of these 100 areal units are fixed, the values of $\boldsymbol{X}$ are repeated for every 100 elements. In addition, we let the coefficients of $\boldsymbol{X}$ to be $\boldsymbol{\beta} = (\beta_1, \beta_2)' = (1, 1)'$.

The first 100 values of $\boldsymbol{X}\boldsymbol{\beta}$ (i.e., values of all 100 spatial locations at time 1) are shown in space in Fig. 4.3. For example, the first areal unit at time 1 is located

at the lower left corner of the panel, with value $\left[x_{11,1}, x_{21,1}\right]\boldsymbol{\beta} = [0.1, 0.1]\,(1,1)' = 0.2$; the second areal unit at time 1 is right next to the previous pixel with value $\left[x_{12,1}, x_{22,1}\right]\boldsymbol{\beta} = [0.2, 0.1]\,(1,1)' = 0.3$; and the 100^{th} areal unit at time 1 is located on the upper right corner with value $\left[x_{1100,1}, x_{2100,1}\right]\boldsymbol{\beta} = [1,1]\,(1,1)' = 2$. It can be seen that the values of $\boldsymbol{X\beta}$ increase toward east and north.

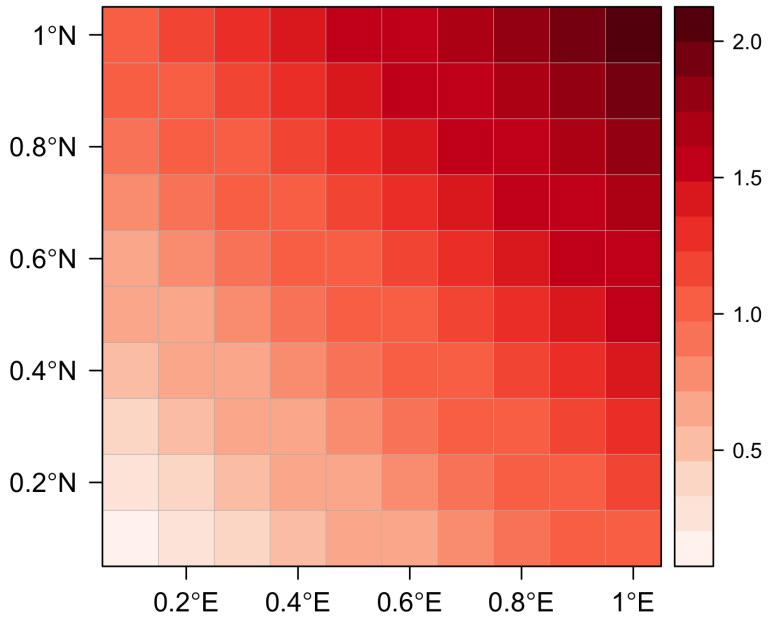

Figure 4.3: The spatial distribution of $\boldsymbol{X\beta}$ at one time point.

Based on the two different $\boldsymbol{MCM}$ as discussed in Section 4.2.1, eigenvalues and eigenvectors of $\boldsymbol{M}_{(1)}\boldsymbol{C}_{ST}\boldsymbol{M}_{(1)}$ and $\boldsymbol{M}_{(X)}\boldsymbol{C}_{ST}\boldsymbol{M}_{(X)}$ are generated. Fig. 4.4 shows the eigenvalues for the $10 \times 10 \times 20$ spatio-temporal domain derived from two alternative $\boldsymbol{MCM}$. In general, the eigenvalues are very similar between these two $\boldsymbol{MCM}$s; 876 out of 2000 eigenvalues are positive for $\boldsymbol{M}_{(1)}\boldsymbol{C}_{ST}\boldsymbol{M}_{(1)}$ (Fig. 4.4(a)), whereas 875 are

positive for $\boldsymbol{M}_{(X)}\boldsymbol{C}_{ST}\boldsymbol{M}_{(X)}$ (Fig. 4.4(b)). The smallest positive eigenvalues in these two cases are both equal to 0.0071 approximately.

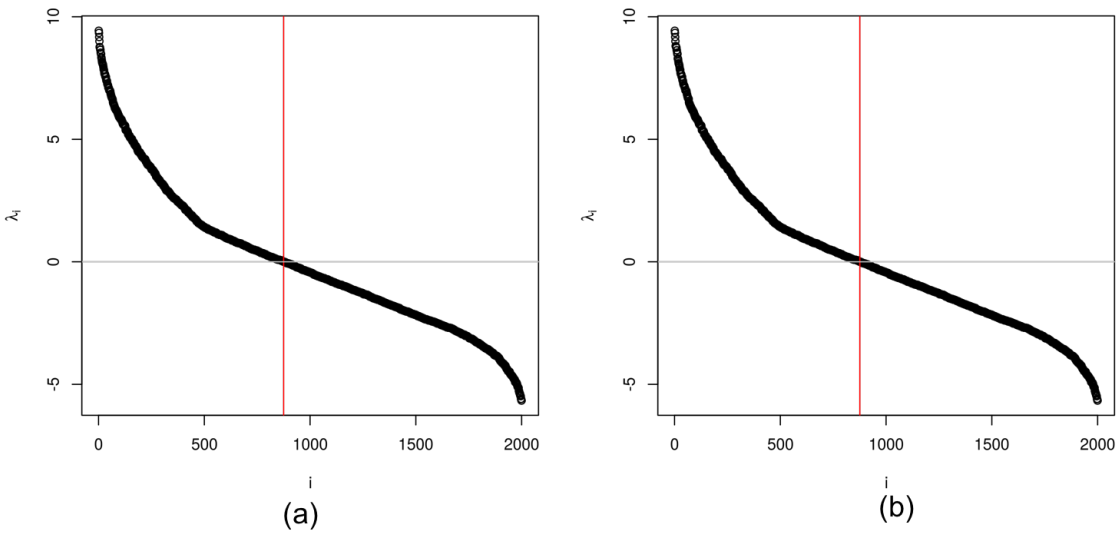

Figure 4.4: Eigenvalues of simulated spatio-temporal domain based on (a) $\boldsymbol{M}_{(1)}\boldsymbol{C}\boldsymbol{M}_{(1)}$ and (b) $\boldsymbol{M}_{(X)}\boldsymbol{C}\boldsymbol{M}_{(X)}$, respectively. The indices of the eigenvalues (in decreasing order) are shown in x-axis, and the corresponding eigenvalues are displayed in y-axis. Red lines indicate the indices of the eigenvalues with smallest positive value, 876 and 875 for (a) $\boldsymbol{M}_{(1)}\boldsymbol{C}\boldsymbol{M}_{(1)}$ and (b) $\boldsymbol{M}_{(X)}\boldsymbol{C}\boldsymbol{M}_{(X)}$, respectively.

Eigenvectors $\boldsymbol{e}_1$, $\boldsymbol{e}_{400}$, and $\boldsymbol{e}_{800}$ from time 1 to time 3 based on $\boldsymbol{M}_{(1)}\boldsymbol{C}\boldsymbol{M}_{(1)}$ and $\boldsymbol{M}_{(X)}\boldsymbol{C}\boldsymbol{M}_{(X)}$ are displayed in Figure 4.5 (a) and (b), respectively. These three eigenvectors are selected as they are associated with large, median, and small positive eigenvalues, respectively. Values in $\boldsymbol{e}_1$ are more homogeneous, reflecting stronger spatial correlation; whereas values in $\boldsymbol{e}_{800}$ are more heterogeneous, indicating smaller scale of spatial dependence. For one specific spatial unit, the values are consistently positive (or negative) over time, showing a temporal correlation structure.

The simulated data field $\boldsymbol{y} = \left(y_{1,1}, y_{2,1}, \ldots, y_{100,20}\right)$ is produced as a combination of

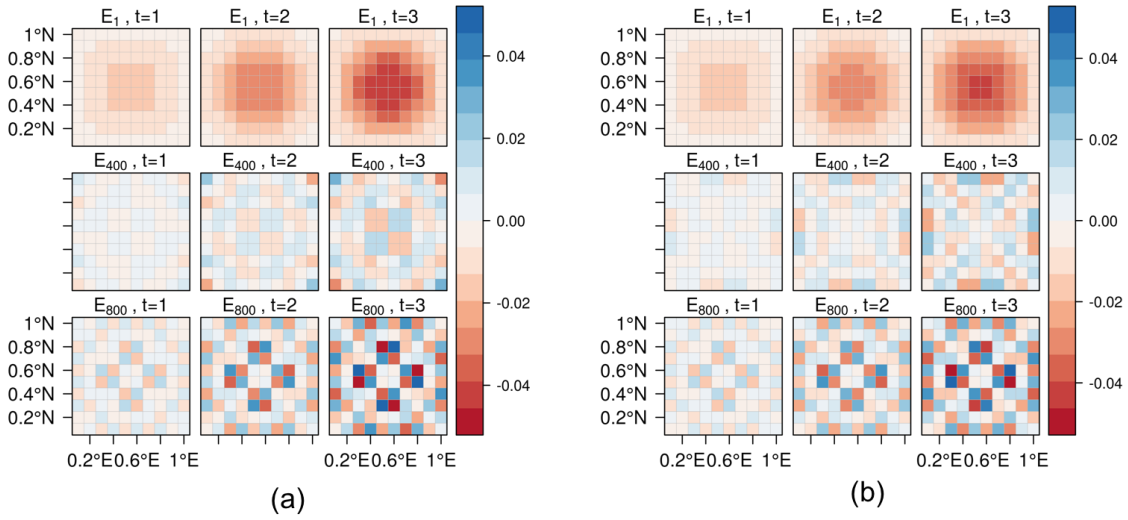

Figure 4.5: Eigenvectors e_1, e_{400}, and e_{800} at time 1, 2, and 3 based on (a) $\boldsymbol{M}_{(1)}\boldsymbol{C}\boldsymbol{M}_{(1)}$ and (b) $\boldsymbol{M}_{(X)}\boldsymbol{C}\boldsymbol{M}_{(X)}$, respectively.

a linear function of covariates $\boldsymbol{X\beta}$ and the space-time mixture structure. To generate $\boldsymbol{y}$, four spatio-temporal autocorrelation structures are implemented. First, two spatio-temporal components are generated in accordance with the STEF mechanism, based on two alternative $\boldsymbol{M}\boldsymbol{C}_{ST}\boldsymbol{M}$. Then, the space-time structure complies with the BHM structures presented in Chapter 3. Finally the spatio-temporal associations are based on a Gaussian random field. Hence, in the last set of experiments the data generating mechanism is not associated with any of the models that are discussed so far. Experiments for these scenarios are named EFM1, EFMX, AR, and RF, respectively; for each scenario 50 replicates are generated.

4.3.2 Scenario 1: Spatio-temporal autocorrelation structure of EFM1

The synthetic response $\boldsymbol{y}$ is generated following Eq. 4.7. $\boldsymbol{M}_{(1)}\boldsymbol{C}_{ST}\boldsymbol{M}_{(1)}$ is used for creating proxy variables (i.e., eigenvectors) for the spatio-temporal residual

structure. In particular, these proxy variables are a set of 50 eigenvectors randomly selected from the first 200 positive eigenvectors of $\boldsymbol{M}_{(1)}\boldsymbol{C}_{ST}\boldsymbol{M}_{(1)}$, $\boldsymbol{E} = \{\boldsymbol{e}_{(1)}, \ldots, \boldsymbol{e}_{(50)}\}$. The corresponding coefficients $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{50})$ are generated from $\gamma_i \sim$ discrete U$(1, 10)$, for $i = 1, \ldots, 50$. In order to evaluate the effects of residual variance, σ^2 is assigned to be 0.001, 1, and 10. Therefore, three sets of simulated $\boldsymbol{y}$ are examined. Fig. 4.6(a) showing the simulated data in the first three time periods over all areal units. Note that values of $\boldsymbol{y}$ are strongly correlated in space and time when $\sigma^2 = 0.001$, whereas values of $\boldsymbol{y}$ are more randomly scattered when $\sigma^2 = 10$. Therefore, spatial-temporal autocorrelation dominates dynamics when σ^2 is small.

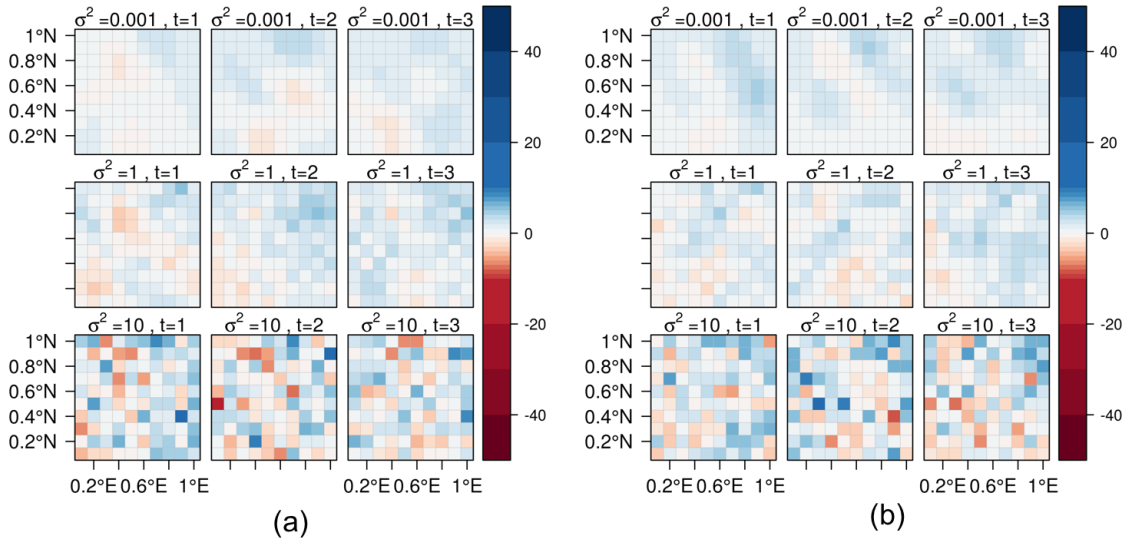

Figure 4.6: Simulated $\boldsymbol{y}$ from $t = 1$ to $t = 3$, for different values of σ^2, using spatio-temporal autocorrelation structure of (a) EFM1 and (b) EFMX, respectively.

4.3.2.1 Scenario 2: Spatio-temporal autocorrelation structure of EFMX

For EFMX, the synthetic response $\boldsymbol{y}$ is generated as in Scenario 1, except that $\boldsymbol{M}_{(X)}\boldsymbol{C}_{ST}\boldsymbol{M}_{(X)}$ is used instead of $\boldsymbol{M}_{(1)}\boldsymbol{C}_{ST}\boldsymbol{M}_{(1)}$. Fig. 4.6(b) shows three sets of simulated $\boldsymbol{y}$ in the first three time periods over all areal units, for different values of σ^2. Similar to Fig. 4.6(a), values of $\boldsymbol{y}$ are strongly correlated in space and time when $\sigma^2 = 0.001$, whereas values of $\boldsymbol{y}$ are more randomly scattered when $\sigma^2 = 10$. Therefore, spatial-temporal autocorrelation dominates dynamics when σ^2 is small. Scenario 1 and scenario 2 are in accordance with the STEF_CF and STEF_NCF approaches, respectively. It is interesting to evaluate how STEF_ICF performs in this case.

4.3.2.2 Scenario 3: AR spatio-temporal autocorrelation structure

For Scenario 3, residual structure per pixel is characterized by an AR(1) structure over time; pixels which are located close in space possess similar autocorrelation parameters ϕ. Hence, for i^{th} pixel at time t, the spatio-temporal component is specified as:

$$\boldsymbol{E}_{i,t} = \phi_i \boldsymbol{E}_{i,t-1} + \epsilon,\ \epsilon \sim N(0, \sigma^2),\ i = 1, \ldots, 100,\ t = 1, \ldots, 20. \tag{4.14}$$

where $\phi_i = \dfrac{max(x_{1i}, x_{2i})}{1.2}$. This ensures that temporal processes in the AR(1) model have $\phi_i < 1$ so they satisfy the stationarity condition (see Fig. 4.7); σ^2 takes values 0.001, 1, and 10, respectively, for 3 different white noise levels. The spatio-temporal random components $\boldsymbol{E}$ are added to the fixed effect part (i.e., $\boldsymbol{X\beta}$), resulting in simulated $\boldsymbol{y}$ with three different structures (Fig. 4.8). Similar to Scenario 1, spatial-temporal autocorrelation is more apparent when σ^2 is small.

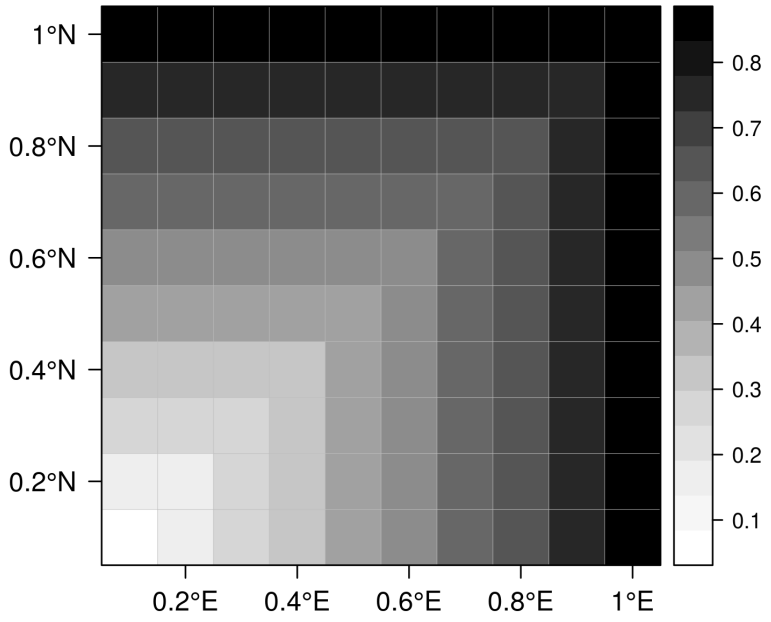

Figure 4.7: The values of ϕ in the spatial domain.

4.3.2.3 Scenario 4: Gaussian random field structure (RF).

For Scenario 4, residuals possess a Gneiting Gaussian random field structure. For a random process $E(\boldsymbol{s},t),(\boldsymbol{s},t) \in \mathbb{R}^2 \times \mathbb{R}$, with the values of a space-time variable indexed at the coordinates $(\boldsymbol{s}_1,t_1),\ldots,(\boldsymbol{s}_{100},t_{20})$, the stationary covariance function of the process is defined as:

$$C(\boldsymbol{h},u) = cov\left(E(\boldsymbol{s},t),E(\boldsymbol{s}+\boldsymbol{h};t+u)\right),(\boldsymbol{h},u) \in \mathbb{R}^2 \times \mathbb{R}, \tag{4.15}$$

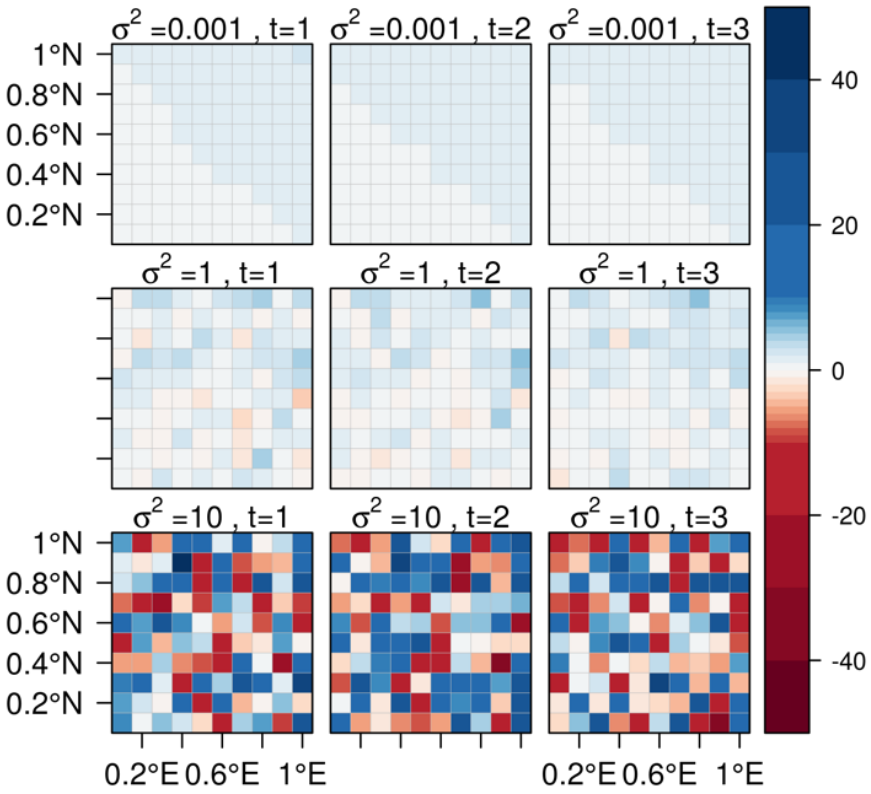

Figure 4.8: Simulated $\boldsymbol{y}$ using AR spatio-temporal autocorrelation structure for different values of σ^2 at $t = 1$, $t = 2$, and $t = 3$, respectively.

where $(\boldsymbol{h}, u)$ is the space-time lag. The Gneiting covariance function is specified as:

$$C(\boldsymbol{h}, u) = \frac{1}{1 + a|u|^{\lambda}} \exp\{-\frac{c\|\boldsymbol{h}\|^{\nu}}{(1 + a|u|^{\lambda})^{0.5\nu\gamma}}\}, \tag{4.16}$$

where a and c are nonnegative temporal and spatial scaling parameters, respectively; λ and ν are temporal and spatial smoothing parameters, respectively, taking values in $[0, 2]$; finally γ is a space-time interaction parameter which takes values in [0,1]: $\gamma = 0$ corresponds to a purely separable model and $\gamma = 1$ to a purely nonseparable model. The values of parameters are chosen to be: $\nu = 1$, $\lambda = 1.544$, $c = 0.00134$, $a = 0.901$,

and $\gamma = 1$ to mimic a valid Gneiting – Gaussian random field process (Gneiting 2002). Fig. 4.9 depicts the generated $\boldsymbol{y}$ from time 1 to time 3.

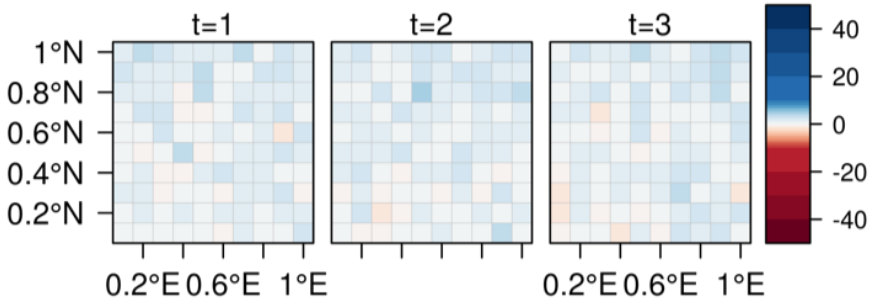

Figure 4.9: Simulated $\boldsymbol{y}$ using RF spatio-temporal autocorrelation structure at $t = 1$, $t = 2$, and $t = 3$, respectively

4.3.3 Experimental design of Monte Carlo simulations

Monte Carlo experiments are performed using multiple STEF specifications combined with alternative modeling strategies as described in section 4.2.2. In order to examine the capability of recovering the true coefficients and the true spatio-temporal structure, STEFs are evaluated on four types of spatio-temporal datasets as described in section 4.3.1. Finally, Bayesian hierarchical spatio-temporal statistical models (i.e., STCARs, including STCARlinear, STCARanova, and STCARar, introduced in Chapter 3) are also applied for comparison. To summarize, 90 Monte Carlo experiments are discussed in what follows: 27 of the experiments are performed using data with EFM1 structure, and 27 of the experiments are performed using data with EFMX structure (Table 4.1); 27 of the experiments are performed on simulated data with AR structure, and 9 experiments analyze data with RF structure (Table 4.2).

Table 4.1: The design of Monte Carlo experiments (part 1)

Monte Carlo experiment	**Spatio-temporal autocorrelation structure of data**	σ^2	**Modeling method**	$\boldsymbol{M}$ **matrix**	**Eigenvectors used for modeling**
EFM1_1 EFM1_2 EFM1_3	EFM1	0.001 1 10	STEF_CF_method1	$\boldsymbol{M}_{(1)}$	positive
EFM1_4 EFM1_5 EFM1_6	EFM1	0.001 1 10	STEF_CF_method2	$\boldsymbol{M}_{(1)}$	positive
EFM1_7 EFM1_8 EFM1_9	EFM1	0.001 1 10	STEF_NCF_method1	$\boldsymbol{M}_{(X)}$	positive
EFM1_10 EFM1_11 EFM1_12	EFM1	0.001 1 10	STEF_NCF_method2	$\boldsymbol{M}_{(X)}$	positive
EFM1_13 EFM1_14 EFM1_15	EFM1	0.001 1 10	STEF_ICF_method1	$\boldsymbol{M}_{(1)}$	all
EFM1_16 EFM1_17 EFM1_18	EFM1	0.001 1 10	STEF_ICF_method2	$\boldsymbol{M}_{(1)}$	all
EFM1_19 EFM1_20 EFM1_21	EFM1	0.001 1 10	STCARlinear		
EFM1_22 EFM1_23 EFM1_24	EFM1	0.001 1 10	STCARanova		
EFM1_25 EFM1_26 EFM1_27	EFM1	0.001 1 10	STCARar		
EFMX_1 EFMX_2 EFMX_3	EFMX	0.001 1 10	STEF_CF_method1	$\boldsymbol{M}_{(1)}$	positive
EFMX_4 EFMX_5 EFMX_6	EFMX	0.001 1 10	STEF_CF_method2	$\boldsymbol{M}_{(1)}$	positive
EFMX_7 EFMX_8 EFMX_9	EFMX	0.001 1 10	STEF_NCF_method1	$\boldsymbol{M}_{(X)}$	positive
EFMX_10 EFMX_11 EFMX_12	EFMX	0.001 1 10	STEF_NCF_method2	$\boldsymbol{M}_{(X)}$	positive
EFMX_13 EFMX_14 EFMX_15	EFMX	0.001 1 10	STEF_ICF_method1	$\boldsymbol{M}_{(1)}$	all
EFMX_16 EFMX_17 EFMX_18	EFMX	0.001 1 10	STEF_ICF_method2	$\boldsymbol{M}_{(1)}$	all
EFMX_19 EFMX_20 EFMX_21	EFMX	0.001 1 10	STCARlinear		
EFMX_22 EFMX_23 EFMX_24	EFMX	0.001 1 10	STCARanova		
EFMX_25 EFMX_26 EFMX_27	EFMX	0.001 1 10	STCARar		

Table 4.2: The design of Monte Carlo experiments (part 2)

Monte Carlo experiment	**Spatio-temporal autocorrelation structure of data**	σ^2	**Modeling method**	$\boldsymbol{M}$ **matrix**	**Eigenvectors used for modeling**
AR_1 AR_2 AR_3	AR	0.001 1 10	STEF_CF_method1	$\boldsymbol{M}_{(1)}$	positive
AR_4 AR_5 AR_6	AR	0.001 1 10	STEF_CF_method2	$\boldsymbol{M}_{(1)}$	positive
AR_7 AR_8 AR_9	AR	0.001 1 10	STEF_NCF_method1	$\boldsymbol{M}_{(X)}$	positive
AR_10 AR_11 AR_12	AR	0.001 1 10	STEF_NCF_method2	$\boldsymbol{M}_{(X)}$	positive
AR_13 AR_14 AR_15	AR	0.001 1 10	STEF_ICF_method1	$\boldsymbol{M}_{(1)}$	all
AR_16 AR_17 AR_18	AR	0.001 1 10	STEF_ICF_method2	$\boldsymbol{M}_{(1)}$	all
AR_19 AR_20 AR_21	AR	0.001 1 10	STCARlinear		
AR_22 AR_23 AR_24	AR	0.001 1 10	STCARanova		
AR_25 AR_26 AR_27	AR	0.001 1 10	STCARar		
RF_1	RF		STEF_CF_method1	$\boldsymbol{M}_{(1)}$	positive
RF_2	RF		STEF_CF_method2	$\boldsymbol{M}_{(1)}$	positive
RF_3	RF		STEF_NCF_method1	$\boldsymbol{M}_{(X)}$	positive
RF_4	RF		STEF_NCF_method2	$\boldsymbol{M}_{(X)}$	positive
RF_5	RF		STEF_ICF_method1	$\boldsymbol{M}_{(1)}$	all
RF_6	RF		STEF_ICF_method2	$\boldsymbol{M}_{(1)}$	all
RF_7	RF		STCARlinear		
RF_8	RF		STCARanova		
RF_9	RF		STCARar		

To evaluate the Monte Carlo experiments, the performance of STEF ensemble members and three STCAR models will be examined. Results are evaluated in terms of eigenvector selection, confidence intervals and their widths, coverage with respect to the true coefficient, as well as computational times. In addition, RMSE and MAE values are calculated to quantify the accuracy (i.e., the distance of estimates $\hat{\beta}_{1_j}$ and $\hat{\beta}_{2_j}$, $j = 1, \ldots, 50$, from the true values $\beta_1 = 1$ and $\beta_2 = 1$, respectively):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{50}(\hat{\beta}_{i_j} - 1)^2}\ , \text{and} \tag{4.17}$$

$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{50} \mid (\hat{\beta}_{i_j} - 1) \mid \tag{4.18}$$

where $i = 1, 2$, and 50 is the total number of replicates for one Monte Carlo experiment.

It is worth noting that all STEF methods are essentially frequentist approaches, whereas STCARs are Hierarchical Bayesian specifications. Hence the uncertainty associated with parameter estimates is illustrated using conventional frequentist confidence intervals (CI) for STEF-based approaches whereas credible intervals (CI) are reported for STCARs. Fundamentally, confidence interval treat their bounds as random variables and the parameter as fixed; whereas credible intervals consider their bounds as fixed and the estimated parameter as a random variable, with knowledge of prior distribution. These two types of CIs differ philosophically, but are still analogous to each other. Therefore, CIs are conducted and compared in this study.

When doing STCAR modeling, different settings with respect to MCMC are specified, based on the complexities of model structure. STCARlinear and STCARanova both use a single chain with a total of 100,000 iterations. The number of iterations for the burnin period is 10,000, and the thinning rate is 100 so the number of samples used for the estimation of the parameters is 900. For STCARar, a single chain with

a total of 500,000 iterations is conducted. The number of iterations for the burnin period is 100,000, and the thinning rate is 100 so the number of samples used for the estimation of the parameters is 4,000. MCMC are evaluated using the Geweke diagnostic (Geweke *et al.* 1991), which should be within the range of -2 and 2. The minimum acceptable effective sample size is set to be equal to 150. According to the diagnostics of convergence, the results of Monte Carlo experiments that follow neglect STCAR estimates which do not appear to have converged after large threshold of MCMC iterations has been exceeded.

4.3.4 Results of Monte Carlo experiments

4.3.4.1 Scenario EFM1

STEF results in different eigenvector selection depending on the selected implementation. The numbers of nonzero eigenvectors selected for scenario EFM1 are shown in Fig. 4.10. As σ^2 increases, the numbers of selected eigenvectors decrease dramatically for both CF and NCF (in this case, signal to noise ratio decreases, so it is harder to identify the correct structure). However, for ICF the number of selected eigenvectors when σ^2 is moderate is significantly larger compared to the corresponding number when σ^2 is large. Using STEF_CF_method1 and STEF_CF_method2, the numbers of selected eigenvectors are 55, 45, and 5 on average for $\sigma^2 = 0.001$, 1, and 10, respectively. Using STEF_NCF_method1 and STEF_NCF_method2, the numbers of selected eigenvectors are similar to CF for $\sigma^2 = 1$, and 10, whereas more eigenvectors (i.e., around 100) are selected when $\sigma^2 = 0.001$. The numbers of selected eigenvectors stay less than 20 for data with extreme σ^2 values when using STEF_ICF,

different from the case of $\sigma^2 = 1$ (i.e. large number of selected eigenvectors). The two methods of two-step STEF specifications (i.e., method1 and method2, described in Section 4.2.2.3) show general agreement in terms of the number of selected eigenvectors, except for a dramatically different performance of CF when σ^2 equals 0.001 (i.e., concentrated around 50 for method1 whereas it ranges from 50 to 100 for method2). In general, STEF_CF performs better in terms of capturing numbers of eigenvectors. This finding is expected a priori, as STEF_CF complies with the data generating mechanism for this scenario.

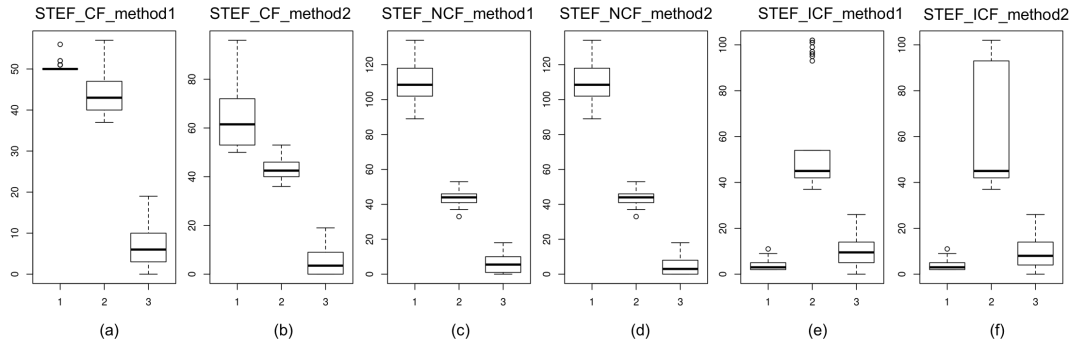

Figure 4.10: Box plots o thef numbers of nonzero eigenvalues selected for scenario EFM1 by(a) STEF_CF_method1, (b) STEF_CF_method2, (c) STEF_NCF_method1, (d) STEF_NCF_method2, (e) STEF_ICF_method1, and (f) STEF_ICF_method2. The numbers 1, 2, and 3 on x-axis represent data with σ^2 of 0.001, 1, and 10, respectively.

The confusion matrix of eigenvector selection for Scenario EFM1 is shown in Table 4.3. STEF_CF captures the majority of eigenvectors that participate in the data generating mechanism when σ^2 is small (i.e. large signal to noise ratio). As the signal to noise ratio drops, STEF excludes eigenvectors that contribute to the data generating mechanism. STEF_NCF performs the worst as the FN, and FP magnitudes are all relatively high. When the random error component is weak, STEF_ICF show good performances with zero FN. Note that the total number of eigenvectors is large

(2000) for STEF_ICF, whereas it is close to 850 for STEF_CF and STEF_NCF. Therefore the number of TN is much larger for STEF_ICF relative to STEF_CF and STEF_NCF, as expected.

Table 4.3: Confusion matrix of eigenvector selection for Scenario EFM1. The numbers of True negative (TN), False negative (FN), False positive (FP), and True positive (TP) are included. Each column is the corresponding mean or standard deviation across 50 replicates. Results corresponding to simulated data with σ^2 equal to 0.001, 1, and 10 are shown in red, black, and blue, respectively. Table cells are colored for displaying results based on different methods.

MC experiments	$\sigma^\star$	Modeling method	Number of eigenvector	TN		FN		FP		TP	
				Mean	sd	Mean	sd	Mean	sd	Mean	sd
EFM1_1	0.001			825.62	0.97	0.00	0.00	0.38	0.97	50.00	0.00
EFM1_2	1	STEF_CF_method1		820.06	3.73	11.88	2.19	5.94	3.73	38.12	2.19
EFM1_3	10		876	824.88	1.39	44.66	3.79	1.12	1.39	5.34	3.79
EFM1_4	0.001			812.18	11.96	0.00	0.00	13.82	11.96	50.00	0.00
EFM1_5	1	STEF_CF_method2		821.28	2.48	11.88	2.12	4.72	2.48	38.12	2.12
EFM1_6	10			825.42	1.03	45.76	4.07	0.58	1.03	4.24	4.07
EFM1_7	0.001			747.68	10.75	17.16	3.16	77.32	10.75	32.84	3.16
EFM1_8	1	STEF_NCF_method1		797.96	3.72	33.38	2.60	27.04	3.72	16.62	2.60
EFM1_9	10		875	820.98	4.01	47.98	2.03	4.02	4.01	2.02	2.03
EFM1_10	0.001			747.68	10.75	17.16	3.16	77.32	10.75	32.84	3.16
EFM1_11	1	STEF_NCF_method2		797.96	3.72	33.38	2.60	27.04	3.72	16.62	2.60
EFM1_12	10			821.86	3.78	48.32	2.02	3.14	3.78	1.68	2.02
EFM1_13	0.001			1950.00	0.00	46.16	2.01	0.00	0.00	3.84	2.01
EFM1_14	1	STEF_ICF_method1		1932.00	21.61	11.10	2.94	18.00	21.61	38.90	2.94
EFM1_15	10		2000	1947.48	2.90	42.38	4.67	2.52	2.90	7.62	4.67
EFM1_16	0.001			1950.00	0.00	46.16	2.01	0.00	0.00	3.84	2.01
EFM1_17	1	STEF_ICF_method2		1930.16	22.33	11.02	2.92	19.84	22.33	38.98	2.92
EFM1_18	10			1947.78	2.70	42.74	4.58	2.22	2.70	7.26	4.58

Among the results of STEFs, STEF_CF performs the best whereas STEF_NCF performs the worst, in terms of accuracy and coverage. The outperformance of STEF_CF can be expected as the data generating mechanism is based on STEF_CF. In addition, the results of STEF_ICF are relatively close to STEF_CF. The medians of $\hat{\beta}_1$ across 50 replicates for all STEF methods, for varying signal to noise rations, are all close to 1 (Tables 4.4). However, STEF_CF and STEF_NCF provide narrower confidence intervals (CI) than STEF_ICF; the CI get wider as the random error of data increases (Tables 4.5). When the σ^2 is low (hence the spatiotemporal component dominates), STEF_CF achieves the best coverage (greater than 90%), whereas STEF_NCF leads to much lower coverage (less than 4%). For realistic values of σ^2, however, STEF_ICF dominates among STEF. In addition, coverage stays close to 85% for all STEF methodologies when σ^2 is large.

With regard to accuracy, RMSE and MAE of STEF increase as σ^2 increases, in accordance with prior expectations. Among the alternative STEF approaches, STEF_CF results in the smallest RMSE and MAE, and STEF_ICF shows relatively smaller RMSE and MAE, compared to STEF_NCF (Table 4.6). However, STEF_ICF dominates for realistic values of σ^2. Overall, the precision and coverage of the $\boldsymbol{\beta}$ by STEF_CF and STEF_ICF do not change significantly as the variability of the random error increases; on the other hand the widths of confidence intervals increase. Despite all the similarities and discrepancies, the two strategies of 2-stage STEF, namely method1 and method2, do not show significant differences.

STCAR estimates are also close to the true values with credible intervals significantly wider relative to STEF (the widest intervals are observed for STCARar whereas the narrowest for STCARlinear). Wide credible intervals lead to higher coverage of the estimated coefficients, as expected (75 to 100%). In terms of RMSE and MAE,

STCAR performance is close to STEF when moderate to weak signal to noise ratios, when signal to noise ratios are moderate or weak; STEF on the other hand is more accurate when the spatio-temporal structure dominates.

STEF_ICF requires shorter time for modeling due to the estimation using smaller number of eigenvector selected based on extra steps of VIF and SIS (Table 4.7). Computational times increase to around 40s, 150s, 120s for STEF_CF or STEF_NCF, STCARlinear, and STCARanova respectively. In addition, much longer time (i.e., 660s) is needed For STCARar. It is worth noting that there exists a MCMC convergence problem for the STCAR models as discussed in Chapter 3. By closely examining convergence through Geweke diagnostics, only a half to two thirds of simulations appear as convergent. In such cases parameter estimation is unreliable; hence the corresponding results are not reported (Table 4.14).

Table 4.4: Parameter estimation of Monte Carlo experiments for Scenario EFM1. Each column is the corresponding median across 50 replicates. Results corresponding to simulated data with σ^2 equal to 0.001, 1, and 10 are shown in red, black, and blue, respectively. Table cells are colored for displaying results based on different methods.

MC experiments	σ^2	Modeling method	β_1			β_2		
			estimate	95%CI		estimate	95%CI	
EFM1_1	0.001		1.0005	0.9960	1.0045	0.9998	0.9955	1.0040
EFM1_2	1	STEF_CF_method1	0.9955	0.8509	1.1596	0.9998	0.8550	1.1451
EFM1_3	10		0.9603	0.5536	1.3783	1.0528	0.6565	1.4333
EFM1_4	0.001		0.9993	0.9884	1.0099	1.0007	0.9902	1.0109
EFM1_5	1	STEF_CF_method2	1.0111	0.8651	1.1332	1.0072	0.8837	1.1324
EFM1_6	10		1.0347	0.6591	1.4090	1.0323	0.6603	1.4071
EFM1_7	0.001		0.9713	0.9677	0.9750	1.0170	1.0133	1.0207
EFM1_8	1	STEF_NCF_method1	1.0150	0.9018	1.1283	1.0119	0.8976	1.1263
EFM1_9	10		1.0305	0.6582	1.3999	1.0503	0.6785	1.4274
EFM1_10	0.001		0.9713	0.9677	0.9750	1.0170	1.0133	1.0207
EFM1_11	1	STEF_NCF_method2	1.0150	0.9018	1.1283	1.0119	0.8976	1.1263
EFM1_12	10		1.0305	0.6572	1.4019	1.0503	0.6785	1.4274
EFM1_13	0.001		0.9560	0.8546	1.0611	1.0148	0.9133	1.1208
EFM1_14	1	STEF_ICF_method1	0.9961	0.8681	1.1275	1.0051	0.8738	1.1284
EFM1_15	10		1.0112	0.6386	1.3938	1.0205	0.6430	1.4047
EFM1_16	0.001		0.9560	0.8546	1.0611	1.0148	0.9133	1.1208
EFM1_17	1	STEF_ICF_method2	1.0122	0.8775	1.1285	0.9956	0.8699	1.1212
EFM1_18	10		1.0344	0.6672	1.4024	1.0205	0.6431	1.3910
EFM1_19	0.001		1.0601	0.7232	1.5048	1.0601	0.5493	1.3861
EFM1_20	1	STCARlinear	1.0770	0.7119	1.3431	1.0770	0.8389	1.4264
EFM1_21	10		1.0200	0.5082	1.5336	1.0200	0.5033	1.5268
EFM1_22	0.001		0.9390	0.6407	1.3775	0.9843	0.6214	1.3352
EFM1_23	1	STCARanova	1.1238	0.7711	1.3716	0.9377	0.6737	1.2394
EFM1_24	10		1.0885	0.5865	1.6183	1.0163	0.4918	1.5306
EFM1_25	0.001		0.9216	0.3906	1.4854	0.9216	0.6090	1.4629
EFM1_26	1	STCARar	1.0833	0.6000	1.6190	1.0833	0.5661	1.6001
EFM1_27	10		0.9972	0.2332	1.6297	0.9972	0.2382	1.6137

Table 4.5: CI length and coverage for Scenario EFM1. Results corresponding to simulated data with σ^2 equal to 0.001, 1, and 10 are shown in red, black, and blue, respectively. Table cells are colored for displaying results based on different methods.

MC experiments	σ^2	Modeling method	CI length of β_1		CI length of β_2		coverage	
			median	Sd	median	Sd	β_1	β_1
EFM1_1	0.001		0.0076	0.0014	0.0076	0.0014	94.00%	94.00%
EFM1_2	1	STEF_CF_method1	0.3370	0.1268	0.3374	0.1270	68.00%	72.00%
EFM1_3	10		0.7647	0.2006	0.7646	0.2009	82.00%	82.00%
EFM1_4	0.001		0.0283	0.0177	0.0283	0.0177	92.00%	90.00%
EFM1_5	1	STEF_CF_method2	0.2362	0.0678	0.2362	0.0679	78.00%	80.00%
EFM1_6	10		0.7459	0.0126	0.7459	0.0126	88.00%	86.00%
EFM1_7	0.001		0.0073	0.0002	0.0073	0.0002	2.00%	4.00%
EFM1_8	1	STEF_NCF_method1	0.2298	0.0038	0.2298	0.0038	48.00%	52.00%
EFM1_9	10		0.7428	0.0148	0.7428	0.0148	88.00%	88.00%
EFM1_10	0.001		0.0073	0.0002	0.0073	0.0002	2.00%	4.00%
EFM1_11	1	STEF_NCF_method2	0.2298	0.0038	0.2298	0.0038	48.00%	52.00%
EFM1_12	10		0.7465	0.0150	0.7465	0.0150	88.00%	88.00%
EFM1_13	0.001		0.2061	0.0186	0.2061	0.0186	66.00%	66.00%
EFM1_14	1	STEF_ICF_method1	0.2383	0.0534	0.2381	0.0533	90.00%	86.00%
EFM1_15	10		0.7437	0.1317	0.7436	0.1306	88.00%	84.00%
EFM1_16	0.001		0.2061	0.0186	0.2061	0.0186	66.00%	66.00%
EFM1_17	1	STEF_ICF_method2	0.2359	0.0460	0.2359	0.0459	82.00%	78.00%
EFM1_18	10		0.7415	0.0150	0.7415	0.0150	88.00%	84.00%
EFM1_19	0.001		0.7795	0.2424	0.7509	0.2482	78.57%	85.71%
EFM1_20	1	STCARlinear	0.5328	0.1935	0.5278	0.1985	86.67%	73.33%
EFM1_21	10		1.0148	0.0601	1.0459	0.0543	93.75%	93.75%
EFM1_22	0.001		0.8013	0.2969	0.7610	0.3046	84.62%	92.31%
EFM1_23	1	STCARanova	0.5435	0.1912	0.5431	0.2202	83.33%	75.00%
EFM1_24	10		1.0393	0.0393	1.0227	0.0467	83.33%	83.33%
EFM1_25	0.001		1.2373	0.2156	0.8540	0.1223	100.00%	100.00%
EFM1_26	1	STCARar	1.0386	0.0865	1.0668	0.0802	100.00%	90.00%
EFM1_27	10		1.3726	0.2544	1.4181	0.2789	100.00%	100.00%

Table 4.6: RMSE and MAE for Scenario EFM1. Results corresponding to simulated data with σ^2 equal to 0.001, 1, and 10 are shown in red, black, and blue, respectively. Table cells are colored for displaying results based on different methods. (To do: check STEF_NCF two methods)

MC experiments	σ^2	Modeling method	RMSE		MAE	
			β_1	β_2	β_1	β_2
EFM1_1	0.001		0.0021	0.0022	0.0018	0.0018
EFM1_2	1	STEF_CF_method1	0.2613	0.2561	0.1729	0.1639
EFM1_3	10		0.3413	0.3860	0.2634	0.2953
EFM1_4	0.001		0.0117	0.0120	0.0080	0.0082
EFM1_5	1	STEF_CF_method2	0.1496	0.1470	0.0954	0.0920
EFM1_6	10		0.2160	0.2478	0.1667	0.1962
EFM1_7	0.001		0.1386	0.1327	0.1037	0.0999
EFM1_8	1	STEF_NCF_method1	0.1857	0.1828	0.1412	0.1371
EFM1_9	10		0.2106	0.2366	0.1654	0.1933
EFM1_10	0.001		0.1386	0.1327	0.1037	0.0999
EFM1_11	1	STEF_NCF_method2	0.1857	0.1828	0.1412	0.1371
EFM1_12	10		0.2106	0.2366	0.1654	0.1933
EFM1_13	0.001		0.1096	0.1054	0.0865	0.0823
EFM1_14	1	STEF_ICF_method1	0.0792	0.0866	0.0615	0.0659
EFM1_15	10		0.2682	0.2799	0.2145	0.2120
EFM1_16	0.001		0.1096	0.1054	0.0865	0.0823
EFM1_17	1	STEF_ICF_method2	0.1259	0.1308	0.0824	0.0868
EFM1_18	10		0.2262	0.2580	0.1763	0.2016
EFM1_19	0.001		0.2320	0.2320	0.1849	0.1849
EFM1_20	1	STCARlinear	0.1712	0.1712	0.1555	0.1555
EFM1_21	10		0.2268	0.2268	0.1722	0.1722
EFM1_22	0.001		0.2158	0.1951	0.1608	0.1342
EFM1_23	1	STCARanova	0.3382	0.3164	0.2348	0.2492
EFM1_24	10		0.3059	0.3216	0.2521	0.2600
EFM1_25	0.001		0.3009	0.3009	0.2571	0.2571
EFM1_26	1	STCARar	0.1815	0.1815	0.1615	0.1615
EFM1_27	10		0.2909	0.2909	0.2090	0.2090

Table 4.7: Computational times of Monte Carlo experiments for Scenario EFM1. Results corresponding to simulated data with σ^2 equal to 0.001, 1, and 10 are shown in red, black, and blue, respectively. Table cells are colored for displaying results based on different methods. For STCARar, computational times correspond to the time taken to reach the specified limit of MCMC samples.

MC experiments	σ^2	Modeling method	computation time	
			median	sd
EFM1_1	0.001		42.5105	2.0212
EFM1_2	1	STEF_CF_method1	43.5345	1.3962
EFM1_3	10		43.4325	1.4501
EFM1_4	0.001		42.8935	1.4639
EFM1_5	1	STEF_CF_method2	42.7985	1.7588
EFM1_6	10		42.5430	1.5519
EFM1_7	0.001		42.0515	1.9313
EFM1_8	1	STEF_NCF_method1	42.8050	1.4856
EFM1_9	10		42.0040	1.4979
EFM1_10	0.001		41.3315	1.8090
EFM1_11	1	STEF_NCF_method2	42.0895	1.1233
EFM1_12	10		42.8250	1.6336
EFM1_13	0.001		14.3410	0.0638
EFM1_14	1	STEF_ICF_method1	15.0740	0.1006
EFM1_15	10		15.7160	0.2781
EFM1_16	0.001		14.3395	0.0212
EFM1_17	1	STEF_ICF_method2	15.0905	0.1483
EFM1_18	10		15.7240	0.4042
EFM1_19	0.001		160.0530	4.9513
EFM1_20	1	STCARlinear	159.3945	1.0391
EFM1_21	10		159.3785	0.6456
EFM1_22	0.001		122.0965	0.6476
EFM1_23	1	STCARanova	122.3565	0.5656
EFM1_24	10		123.6325	1.5100
EFM1_25	0.001		667.1880	8.9738
EFM1_26	1	STCARar	668.1145	2.3917
EFM1_27	10		680.4655	8.5840

Table 4.8: Percentage of convergent replications of models based on STCAR for scenario EF. Models with Geweke diagnostics values within -2 and 2 are considered as convergent.

	$\sigma^2 = 0.001$	$\sigma^2 = 1$	$\sigma^2 = 10$
STCARlinear	0.70	0.75	0.80
STCARanova	0.65	0.60	0.60
STCARar	0.20	0.50	0.25

4.3.4.2 Scenario EFMX

The average number of selected eigenvectors ranges from 5 to 100, from 5 to 50, and from 5 to 25 for STEF_CF, STEF_NCF, and STEF_ICF, respectively (Fig. 4.11). As the variability for the random error component increases, the average numbers of selected eigenvectors decrease for all STEF approaches. When the data generating mechanism is based on STEF_NCF and σ^2 is small, STEF_CF selects less eigenvectors than STEF_NCF. The number is further reduced for STEF_ICF through the VIF-SIS steps.

The confusion matrix of eigenvector selection for Scenario EFMX is shown in Table 4.9. As random error variability increases, true and false negatives increase, whereas false and true positives decrease for all STEFs. However, STEF_NCF captures the most eigenvectors compared to STEF_CF and STEF_ICF. Specifically for large signal to noise ratios, STEF_NCF identifies all 50 eigenvectors that contribute to the data generating mechanism without any false negatives.

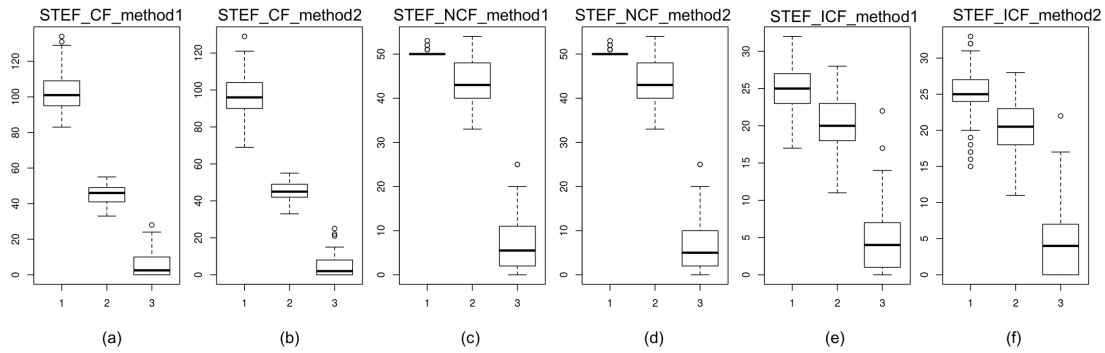

Figure 4.11: Box plots of numbers of nonzero eigenvalues selected for scenario EFMX by(a) STEF_CF_method1, (b) STEF_CF_method2, (c) STEF_NCF_method1, (d) STEF_NCF_method2, (e) STEF_ICF_method1, and (f) STEF_ICF_method2. The numbers 1, 2, and 3 on x-axis represent data with σ^2 of 0.001, 1, and 10, respectively.

Table 4.9: Confusion matrix of eigenvector selection for Scenario EFMX. Each column is the corresponding mean or standard deviation across 50 replicates. Results corresponding to simulated data with σ^2 equal to 0.001, 1, and 10 are shown in red, black, and blue, respectively. Table cells are colored for displaying results based on different methods.

MC experiments	$\sigma^\star$	Modeling method	Number of eigenvector	TN		FN		FP		TP	
				Mean	sd	Mean	sd	Mean	sd	Mean	sd
EFMX_1	0.001	STEF_CF_method1	876	755.16	11.34	18.02	2.74	70.84	11.34	31.98	2.74
EFMX_2	1			798.10	4.36	32.14	2.76	27.90	4.36	17.86	2.76
EFMX_3	10			822.48	4.63	48.06	2.20	3.52	4.63	1.94	2.20
EFMX_4	0.001	STEF_CF_method2		760.60	10.78	18.70	3.09	65.40	10.78	31.30	3.09
EFMX_5	1			798.44	4.44	32.28	2.68	27.56	4.44	17.72	2.68
EFMX_6	10			823.14	4.16	48.18	2.34	2.86	4.16	1.82	2.34
EFMX_7	0.001	STEF_NCF_method1	875	824.66	0.72	0.00	0.00	0.34	0.72	50.00	0.00
EFMX_8	1			819.88	3.61	11.54	2.14	5.12	3.61	38.46	2.14
EFMX_9	10			823.88	1.90	43.90	5.02	1.12	1.90	6.10	5.02
EFMX_10	0.001	STEF_NCF_method2		824.40	0.44	0.00	0.00	0.76	0.47	50.00	0.00
EFMX_11	1			818.53	3.55	12.19	3.64	4.66	4.59	38.83	2.98
EFMX_12	10			823.83	2.43	45.34	6.44	0.25	2.07	4.82	4.24
EFMX_13	0.001	STEF_ICF_method1	2000	1934.50	3.36	40.38	2.60	15.50	3.36	9.62	2.60
EFMX_14	1			1937.94	2.77	41.90	2.31	12.06	2.77	8.10	2.31
EFMX_15	10			1946.78	4.08	48.58	1.62	3.22	4.08	1.42	1.62
EFMX_16	0.001	STEF_ICF_method2		1934.64	3.48	40.32	2.53	15.36	3.48	9.68	2.53
EFMX_17	1			1937.94	2.77	41.90	2.34	12.06	2.77	8.10	2.34
EFMX_18	10			1946.84	4.09	48.60	1.64	3.16	4.09	1.40	1.64

For scenario EFMX, data are characterized by the EF spatio-temporal correlation structure with eigenvectors from $\boldsymbol{M}_X$ (Tables 4.10 to 4.13), thus STEF_NCF outperforms other STEF methods in terms of coverage and accuracy. The median value of $\hat{\beta}_1$ and $\hat{\beta}_2$ are close to 1 for both STEFs and STCARs (Tables 4.10). CI width ranges from 0.01 to 0.8 when derived using STEFs, with STEF_NCF the smallest and STEF_ICF the largest (Tables 4.11). In accordance with prior expectations, CI width increases with σ^2.

All estimated coverages are higher than 70% for STEF methods, except when STEF_CF is used with large signal to noise ratios: : in this case the estimated coverage is less than 30% (Table 4.11). The lowest values of RMSE and MAE are achieved by STEF_NCF, in accordance with a-priori expectations; the worst performance is observed for STEF_CF (Table 4.12). STEF_ICF is close to the best performing STEF_NCF for realistic values of σ^2. Computational times are around 10s shorter for STEF_NCF relative to STEF_CF. In addition, STEF_ICF requires the shortest time for modeling. The fast computation of STEF_ICF is because its VIF-SIS step removes more eigenvectors through screening, accelerating the eigenvector selection and model fitting (Tables 4.13). Despite these variability of results, the two approaches of 2-stage STEF (i.e. method1 and method2) are not significantly different.

The confidence intervals are wider for STCAR models, as their width ranges from 0.8 to 1.4. When the random error of data is small or moderate, the coverage may reach 100%; such estimates are based on small sample sizes though, as a significant percentage of STCAR replicates did not converge (Table 4.14). The accuracy of STCARs are lower, with RMSE and MAE higher than 0.5. Nevertheless, STCARlinear is the best performing approach among STCARs; its performance is close to STEF_ICF.

The computational time required for STCARs is about 10 times the time required to implement STEFs.

Table 4.10: Parameter estimation of Monte Carlo experiments for Scenario EFMX. Each value is the corresponding median across 50 replicates. Results corresponding to simulated data with σ^2 equal to 0.001, 1, and 10 are shown in red, black, and blue, respectively. Table cells are colored for displaying results based on different methods.

MC experiments	σ^2	Modeling method	β_1			β_2		
			estimate	95%CI		estimate	95%CI	
EFMX_1	0.001		1.0452	1.0337	1.0552	0.9808	0.9703	0.9932
EFMX_2	1	STEF_CF_method1	0.9875	0.8555	1.1196	1.0237	0.8774	1.1594
EFMX_3	10		1.0041	0.5813	1.4305	0.9783	0.5721	1.3901
EFMX_4	0.001		1.0255	1.0123	1.0382	0.9945	0.9855	1.0054
EFMX_5	1	STEF_CF_method2	0.9938	0.8659	1.1161	1.0116	0.8669	1.1576
EFMX_6	10		0.9588	0.5832	1.3362	1.0339	0.6628	1.4064
EFMX_7	0.001		1.0000	0.9964	1.0037	1.0000	0.9964	1.0036
EFMX_8	1	STEF_NCF_method1	1.0007	0.8848	1.1169	1.0030	0.8870	1.1190
EFMX_9	10		1.0462	0.6778	1.4145	0.9790	0.6121	1.3451
EFMX_10	0.001		1.0000	0.9964	1.0037	1.0000	0.9964	1.0036
EFMX_11	1	STEF_NCF_method2	1.0007	0.8848	1.1169	1.0030	0.8870	1.1190
EFMX_12	10		1.0462	0.6778	1.4145	0.9790	0.6110	1.3451
EFMX_13	0.001		1.0073	0.9299	1.0851	1.0000	0.9212	1.0819
EFMX_14	1	STEF_ICF_method1	0.9850	0.8422	1.1356	1.0086	0.8693	1.1493
EFMX_15	10		0.9563	0.5840	1.3312	1.0237	0.6555	1.3913
EFMX_16	0.001		1.0065	0.9301	1.0858	0.9993	0.9171	1.0821
EFMX_17	1	STEF_ICF_method2	0.9894	0.8433	1.1356	1.0031	0.8629	1.1396
EFMX_18	10		0.9563	0.5840	1.3312	1.0237	0.6555	1.3913
EFMX_19	0.001		1.0289	0.4979	1.5303	1.0289	0.5190	1.5642
EFMX_20	1	STCARlinear	1.0181	0.6411	1.3295	1.0181	0.6956	1.3934
EFMX_21	10		1.0577	0.4744	1.5603	1.0577	0.6034	1.6758
EFMX_22	0.001		1.0462	0.5116	1.4313	1.0449	0.5098	1.5887
EFMX_23	1	STCARanova	1.0513	0.6710	1.4102	1.0384	0.6907	1.4328
EFMX_24	10		1.1327	0.6551	1.6416	1.1384	0.5454	1.6601
EFMX_25	0.001		1.0320	0.4539	1.3950	1.0320	0.3344	1.4022
EFMX_26	1	STCARar	1.0737	0.4895	1.6339	1.0737	0.5462	1.6711
EFMX_27	10		1.1084	0.4765	1.9884	1.1084	0.3830	1.8172

Table 4.11: CI length and coverage for Scenario EFMX. Each value is the corresponding median across 50 replicates. Results corresponding to simulated data with σ^2 equal to 0.001, 1, and 10 are shown in red, black, and blue, respectively. Table cells are colored for displaying results based on different methods.

MC experiments	σ^2	Modeling method	CI length of β_1		CI length of β_2		coverage	
			median	Sd	median	Sd	β_1	β_1
EFMX_1	0.001		0.0241	0.0086	0.0241	0.0086	8.00%	6.00%
EFMX_2	1	STEF_CF_method1	0.2508	0.1472	0.2508	0.1473	80.00%	74.00%
EFMX_3	10		0.7570	0.1955	0.7570	0.1960	92.00%	90.00%
EFMX_4	0.001		0.0294	0.0102	0.0294	0.0102	26.00%	22.00%
EFMX_5	1	STEF_CF_method2	0.2385	0.0904	0.2384	0.0905	86.00%	90.00%
EFMX_6	10		0.7505	0.0145	0.7505	0.0145	98.00%	98.00%
EFMX_7	0.001		0.0072	0.0001	0.0072	0.0001	94.00%	96.00%
EFMX_8	1	STEF_NCF_method1	0.2286	0.0038	0.2286	0.0038	100.00%	98.00%
EFMX_9	10		0.7419	0.0149	0.7419	0.0149	98.00%	100.00%
EFMX_10	0.001		0.0072	0.0001	0.0072	0.0001	94.00%	96.00%
EFMX_11	1	STEF_NCF_method2	0.2286	0.0038	0.2286	0.0038	100.00%	98.00%
EFMX_12	10		0.7437	0.0162	0.7437	0.0162	98.00%	100.00%
EFMX_13	0.001		0.1661	0.0331	0.1660	0.0330	74.00%	86.00%
EFMX_14	1	STEF_ICF_method1	0.2848	0.0209	0.2848	0.0209	94.00%	94.00%
EFMX_15	10		0.7501	0.0398	0.7501	0.0398	98.00%	98.00%
EFMX_16	0.001		0.1661	0.0299	0.1660	0.0300	80.00%	88.00%
EFMX_17	1	STEF_ICF_method2	0.2844	0.0208	0.2843	0.0208	94.00%	94.00%
EFMX_18	10		0.7507	0.0139	0.7507	0.0139	98.00%	98.00%
EFMX_19	0.001		1.0522	0.2768	1.0012	0.2914	100.00%	100.00%
EFMX_20	1	STCARlinear	0.7234	0.2739	0.7430	0.2760	100.00%	100.00%
EFMX_21	10		1.0333	0.0526	0.9984	0.0572	93.33%	93.33%
EFMX_22	0.001		1.0699	0.3433	1.1064	0.3619	100.00%	100.00%
EFMX_23	1	STCARanova	0.8024	0.3466	0.8102	0.3643	100.00%	100.00%
EFMX_24	10		1.0127	0.0542	1.0514	0.0879	90.00%	80.00%
EFMX_25	0.001		1.0295	0.1213	0.9044	0.2776	100.00%	100.00%
EFMX_26	1	STCARar	1.1627	0.1023	1.0974	0.1109	100.00%	100.00%
EFMX_27	10		1.3936	0.3448	1.3998	0.3427	87.50%	87.50%

Table 4.12: RMSE and MAE for Scenario EFMX. Results corresponding to simulated data with σ^2 equal to 0.001, 1, and 10 are shown in red, black, and blue, respectively. Table cells are colored for displaying results based on different methods.

MC experiments	σ^2	Modeling method	RMSE		MAE	
			β_1	β_2	β_1	β_2
EFMX_1	0.001		0.0679	0.0618	0.0550	0.0518
EFMX_2	1	STEF_CF_method1	0.2095	0.2033	0.1389	0.1357
EFMX_3	10		0.2821	0.2766	0.2067	0.2095
EFMX_4	0.001		0.0519	0.0475	0.0397	0.0373
EFMX_5	1	STEF_CF_method2	0.0876	0.0906	0.0663	0.0758
EFMX_6	10		0.1681	0.1565	0.1330	0.1249
EFMX_7	0.001		0.0019	0.0018	0.0015	0.0014
EFMX_8	1	STEF_NCF_method1	0.0521	0.0499	0.0426	0.0425
EFMX_9	10		0.1752	0.1581	0.1455	0.1201
EFMX_10	0.001		0.0019	0.0018	0.0015	0.0014
EFMX_11	1	STEF_NCF_method2	0.0521	0.0499	0.0426	0.0425
EFMX_12	10		0.1752	0.1581	0.1455	0.1201
EFMX_13	0.001		0.0903	0.0871	0.0578	0.0533
EFMX_14	1	STEF_ICF_method1	0.0978	0.0986	0.0670	0.0686
EFMX_15	10		0.1764	0.1596	0.1404	0.1222
EFMX_16	0.001		0.0736	0.0763	0.0491	0.0455
EFMX_17	1	STEF_ICF_method2	0.0961	0.0972	0.0652	0.0670
EFMX_18	10		0.1758	0.1579	0.1395	0.1210
EFMX_19	0.001		0.0719	0.0719	0.0565	0.0565
EFMX_20	1	STCARlinear	0.1160	0.1160	0.0944	0.0944
EFMX_21	10		0.2454	0.2454	0.1781	0.1781
EFMX_22	0.001		0.0977	0.0805	0.0760	0.0612
EFMX_23	1	STCARanova	0.2031	0.1921	0.1595	0.1324
EFMX_24	10		0.2999	0.3138	0.2476	0.2652
EFMX_25	0.001		0.1968	0.1968	0.1812	0.1812
EFMX_26	1	STCARar	0.2147	0.2147	0.1791	0.1791
EFMX_27	10		0.2776	0.2776	0.2116	0.2116

Table 4.13: Computational times of Monte Carlo experiments for Scenario EFMX. Results corresponding to simulated data with σ^2 equal to 0.001, 1, and 10 are shown in red, black, and blue, respectively. Table cells are colored for displaying results based on different methods.

MC experiments	σ^2	Modeling method	computation time	
			median	sd
EFMX_1	0.001		42.7425	1.5440
EFMX_2	1	STEF_CF_method1	42.7440	1.4962
EFMX_3	10		43.0965	1.2726
EFMX_4	0.001		41.6610	1.6117
EFMX_5	1	STEF_CF_method2	42.0165	1.8203
EFMX_6	10		42.5565	1.3642
EFMX_7	0.001		27.9990	2.3704
EFMX_8	1	STEF_NCF_method1	28.2675	3.3698
EFMX_9	10		29.0260	3.3962
EFMX_10	0.001		27.6560	0.5791
EFMX_11	1	STEF_NCF_method2	27.7075	0.6518
EFMX_12	10		27.5960	0.5301
EFMX_13	0.001		14.6780	0.0794
EFMX_14	1	STEF_ICF_method1	14.7950	0.0451
EFMX_15	10		14.8245	0.0327
EFMX_16	0.001		14.6725	0.0526
EFMX_17	1	STEF_ICF_method2	14.7715	0.0455
EFMX_18	10		14.8170	0.0353
EFMX_19	0.001		160.0530	5.3354
EFMX_20	1	STCARlinear	159.3080	1.0932
EFMX_21	10		159.4725	0.5712
EFMX_22	0.001		122.2265	0.6241
EFMX_23	1	STCARanova	122.3960	0.5952
EFMX_24	10		123.6650	1.6923
EFMX_25	0.001		666.9475	10.1538
EFMX_26	1	STCARar	668.4720	2.4693
EFMX_27	10		682.1515	9.5181

It is worth noting that there exists a MCMC convergence problem for the STCAR models as discussed in Chapter 3. By closely examining convergence through Geweke diagnostics, only a half to two thirds of simulations appear as convergent. Parameter estimation and corresponding inference are based on the convergent replicates only (Table 4.14).

Table 4.14: Percentage of convergent replications of models based on STCAR for scenario EF. Models with Geweke diagnostics values within -2 and 2 are considered as convergent

	$\sigma^2 = 0.001$	$\sigma^2 = 1$	$\sigma^2 = 10$
STCARlinear	0.62	0.58	0.5
STCARanova	0.74	0.68	0.64
STCARar	0.72	0.64	0.44

4.3.4.3 Scenario AR

For scenario AR, STEF_CF and STEF_NCF result in similar numbers of nonzero eigenvectors (around 80) for different values of σ^2 (Fig. 4.12). However, around 10 eigenvectors are selected using STEF_ICF for data with small random error, while 40 eigenvectors are selected on average when data have moderate to large random error. The small number of selected eigenvectors by STEF_ICF when σ^2 is small, may due to the exclusion of large number of eigenvectors through SIS screening process.

STEF and STCAR estimates are all close to the true values for small and moderate σ^2 when signal to noise ratio is large or moderate in this scenario, which favors STCARar estimates (Table 4.15). For $\sigma^2 = 10$, however, STEF_CF_method1 seems to perform significantly better than STEF_CF_method2, and the rest of STEFs. In

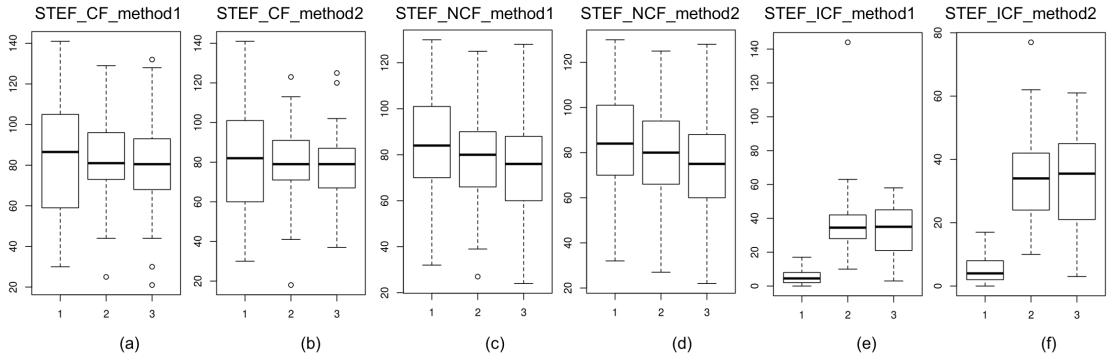

Figure 4.12: Box plots of numbers of nonzero eigenvalues selected for scenario AR by (a) STEF_CF_method1, (b) STEF_CF_method2, (c) STEF_NCF_method1, (d) STEF_NCF_method2, (e) STEF_ICF_method1, and (f) STEF_ICF_method2. The numbers 1, 2, and 3 on x-axis represent data with σ^2 of 0.001, 1, and 10, respectively.

addition, mis-specification of STCAR (i.e., STCARlinear or STCARanova, rather than STCARar) seems to have dramatic consequences.

The average CI width of STEF equal 0.0002, 0.2 and 2 for $\sigma^2 = 0.001, 1, \text{and} 10$, respectively (Table 4.16). CI widths derived from STCAR are wider relative to the ones derived from STEF. Specifically, CI widths derived by STCARlinear and STCARanova are on average equal to 0.03, 1.5, and 13 for $\sigma^2 = 0.001, 1, and 10$ respectively. Average CI widths reduce to 0.03, 0.8 and 6 for STCARar.

STCAR appear to possess high coverage rates, but these rates are not so reliable since they are based on a small number of replications due to convergence issues (Table 4.19). Nevertheless, the coverage of STCAR is superior to that of STEF. In addition, STCARar shows the best performance in terms of RMSE and MAE as one would expect a-priori since it is compliant with the data generating mechanism (Table 4.17). The effects of mis-specification appear to be significant: STCARlinear and STCARanova do not perform well in terms of accuracy. STEF_CF_method1 does not perform well neither with high RMSE and MAE; however, STEF_NCF

performs better and STEF_ICF is close to STEF_NCF, both displaying satisfactory performance.

The computational times for STEF_CF and STEF_NCF are similar, but times decrease dramatically for STEF_ICF which uses VIF and SIS procedures (Table 4.18). As before, computational times for STCAR models are high. The convergence rates for STCAR experiments range from 46% to 76% (Table 4.19). This is to say, STCAR methods have a serious disadvantage if they are so slow to converge even if the data generating mechanism complies with their design. Thus if we take into account accuracy, computational times and convergence rates, STEF methods (especially STEF_ICF) are superior to STCAR.

Table 4.15: Parameter estimation for Scenario AR. Each value is the corresponding median across 50 replicates. Results corresponding to simulated data with σ^2 equal to 0.001, 1, and 10 are shown in red, black, and blue, respectively. Table cells are colored for displaying results based on different methods.

MC experiments	σ^2	Modeling method	β_1			β_2		
			estimate	95%CI		estimate	95%CI	
AR_1	0.001		0.9999	0.9997	1.0001	1.0001	0.9999	1.0004
AR_2	1	STEF_CF_method1	0.9753	0.8127	1.2222	1.0354	0.7951	1.2317
AR_3	10		1.0154	-0.9501	3.0597	1.0682	-0.8591	2.9304
AR_4	0.001		1.0000	0.9999	1.0002	1.0000	0.9999	1.0002
AR_5	1	STEF_CF_method2	1.0230	0.8907	1.1625	0.9718	0.8413	1.1057
AR_6	10		0.4648	-0.9475	1.8069	2.2068	0.8867	3.4815
AR_7	0.001		1.0000	0.9999	1.0001	1.0000	0.9999	1.0001
AR_8	1	STEF_NCF_method1	0.9831	0.8601	1.1061	0.9693	0.8483	1.0887
AR_9	10		0.5092	-0.7141	1.7656	1.7928	0.5414	3.0615
AR_10	0.001		1.0000	0.9999	1.0001	1.0000	0.9999	1.0001
AR_11	1	STEF_NCF_method2	0.9831	0.8601	1.1061	0.9693	0.8483	1.0887
AR_12	10		0.5092	-0.7268	1.7897	1.7928	0.5414	3.1337
AR_13	0.001		1.0000	0.9999	1.0002	1.0000	0.9999	1.0002
AR_14	1	STEF_ICF_method1	0.9432	0.7902	1.0900	0.8827	0.7364	1.0478
AR_15	10		0.5152	-0.9238	1.9173	1.8018	0.3273	3.2600
AR_16	0.001		1.0000	0.9999	1.0002	1.0000	0.9998	1.0002
AR_17	1	STEF_ICF_method2	0.9761	0.8410	1.1188	0.9846	0.8424	1.1228
AR_18	10		0.5152	-0.9238	1.9173	1.8379	0.4229	3.3835
AR_19	0.001		1.0010	0.9826	1.0165	1.0010	0.9826	1.0151
AR_20	1	STCARlinear	0.8568	0.0892	1.6803	0.8568	0.2760	1.7667
AR_21	10		-0.3873	-5.5111	5.8531	-0.3873	-3.9672	9.2695
AR_22	0.001		0.9999	0.9821	1.0160	0.9993	0.9825	1.0166
AR_23	1	STCARanova	0.8216	0.0681	1.5899	1.0479	0.3591	1.7464
AR_24	10		-0.4026	-6.9460	6.4161	2.0049	-5.6105	8.1311
AR_25	0.001		0.9999	0.9981	1.0017	0.9999	0.9986	1.0022
AR_26	1	STCARar	0.9539	0.5171	1.4024	0.9539	0.6730	1.5441
AR_27	10		1.4224	-1.6260	4.1034	1.4224	-4.0857	2.6154

Table 4.16: CI length and coverage for Scenario AR. Each value is the corresponding median across 50 replicates. Results corresponding to simulated data with σ^2 equal to 0.001, 1, and 10 are shown in red, black, and blue, respectively. Table cells are colored for displaying results based on different methods.

MC experiments	σ^2	Modeling method	CI length of β_1		CI length of β_2		coverage	
			median	Sd	median	Sd	β_1	β_2
AR_1	0.001		0.0005	0.0002	0.0005	0.0002	36.00%	26.00%
AR_2	1	STEF_CF_method1	0.4958	0.1922	0.4959	0.1923	16.00%	12.00%
AR_3	10		4.4082	1.4395	4.4149	1.4406	48.00%	40.00%
AR_4	0.001		0.0003	0.0001	0.0003	0.0001	42.00%	42.00%
AR_5	1	STEF_CF_method2	0.2642	0.0579	0.2641	0.0580	46.00%	40.00%
AR_6	10		2.6547	0.6973	2.6549	0.6980	48.00%	48.00%
AR_7	0.001		0.0002	0.0000	0.0002	0.0000	36.00%	42.00%
AR_8	1	STEF_NCF_method1	0.2504	0.0116	0.2504	0.0116	32.00%	30.00%
AR_9	10		2.4810	0.1094	2.4810	0.1094	46.00%	38.00%
AR_10	0.001		0.0002	0.0000	0.0002	0.0000	36.00%	42.00%
AR_11	1	STEF_NCF_method2	0.2504	0.0116	0.2504	0.0116	32.00%	30.00%
AR_12	10		2.4915	0.1140	2.4915	0.1140	46.00%	38.00%
AR_13	0.001		0.0003	0.0000	0.0003	0.0000	30.00%	52.00%
AR_14	1	STEF_ICF_method1	0.2997	0.0688	0.2998	0.0684	44.00%	32.00%
AR_15	10		2.8736	0.6107	2.8733	0.6060	48.00%	52.00%
AR_16	0.001		0.0003	0.0000	0.0003	0.0000	30.00%	54.00%
AR_17	1	STEF_ICF_method2	0.2876	0.0332	0.2876	0.0332	42.00%	36.00%
AR_18	10		2.8675	0.2717	2.8666	0.2722	50.00%	46.00%
AR_19	0.001		0.0342	0.0030	0.0322	0.0036	100.00%	100.00%
AR_20	1	STCARlinear	1.4676	0.2439	1.5048	0.2406	100.00%	100.00%
AR_21	10		12.6832	3.3978	12.2392	3.3941	100.00%	100.00%
AR_22	0.001		0.0331	0.0026	0.0332	0.0033	100.00%	100.00%
AR_23	1	STCARanova	1.4533	0.2178	1.5047	0.2285	92.31%	100.00%
AR_24	10		13.4825	2.9377	13.5397	2.6753	100.00%	100.00%
AR_25	0.001		0.0036	0.0001	0.0036	0.0001	100.00%	100.00%
AR_26	1	STCARar	0.8271	0.0710	0.8478	0.0692	68.75%	87.50%
AR_27	10		6.2662	0.8210	5.4693	2.1809	75.00%	75.00%

Table 4.17: RMSE and MAE for Scenario AR. Results corresponding to simulated data with σ^2 equal to 0.001, 1, and 10 are shown in red, black, and blue, respectively. Table cells are colored for displaying results based on different methods.

MC experiments	σ^2	Modeling method	RMSE		MAE	
			β_1	β_2	β_1	β_2
AR_1	0.001		0.0006	0.0006	0.0004	0.0005
AR_2	1	STEF_CF_method1	0.8081	0.8052	0.6417	0.6467
AR_3	10		3.8819	3.8612	2.9160	3.0364
AR_4	0.001		0.0003	0.0003	0.0002	0.0002
AR_5	1	STEF_CF_method2	0.2308	0.2470	0.1884	0.2027
AR_6	10		2.1576	2.1159	1.6971	1.7996
AR_7	0.001		0.0002	0.0002	0.0002	0.0002
AR_8	1	STEF_NCF_method1	0.2375	0.2579	0.1976	0.2165
AR_9	10		2.1086	2.1544	1.6949	1.7611
AR_10	0.001		0.0002	0.0002	0.0002	0.0002
AR_11	1	STEF_NCF_method2	0.2375	0.2579	0.1976	0.2165
AR_12	10		2.1086	2.1544	1.6949	1.7611
AR_13	0.001		0.0002	0.0002	0.0002	0.0002
AR_14	1	STEF_ICF_method1	0.3367	0.3437	0.2551	0.2715
AR_15	10		2.3645	2.0136	1.7522	1.6528
AR_16	0.001		0.0002	0.0002	0.0002	0.0002
AR_17	1	STEF_ICF_method2	0.2486	0.2735	0.2002	0.2179
AR_18	10		2.0455	2.0329	1.6207	1.6841
AR_19	0.001		0.0012	0.0012	0.0011	0.0011
AR_20	1	STCARlinear	0.2735	0.2735	0.2390	0.2390
AR_21	10		1.7448	1.7448	1.5332	1.5332
AR_22	0.001		0.0012	0.0014	0.0008	0.0012
AR_23	1	STCARanova	0.3204	0.2803	0.2909	0.2267
AR_24	10		2.7691	2.2557	2.4641	1.9560
AR_25	0.001		0.0003	0.0003	0.0002	0.0002
AR_26	1	STCARar	0.3877	0.3877	0.3253	0.3253
AR_27	10		1.6939	1.6939	1.2755	1.2755

Table 4.18: Computational time of Monte Carlo experiment for Scenario AR. Results corresponding to simulated data with σ^2 equal to 0.001, 1, and 10 are shown in red, black, and blue, respectively. Table cells are colored for displaying results based on different methods. For STCARar, computational times correspond to the time taken to reach the specified limit of MCMC samples.

MC experiments	$sigma^2$	Modeling method	computation time	
			median	sd
AR_1	0.001		27.3650	0.7227
AR_2	1	STEF_CF_method1	27.4350	0.6649
AR_3	10		27.5475	0.8406
AR_4	0.001		27.7070	1.0551
AR_5	1	STEF_CF_method2	28.6310	0.7479
AR_6	10		28.9200	0.7168
AR_7	0.001		28.7425	0.9518
AR_8	1	STEF_NCF_method1	28.8800	0.8158
AR_9	10		28.7380	0.6743
AR_10	0.001		28.5210	0.7150
AR_11	1	STEF_NCF_method2	28.3880	0.6447
AR_12	10		28.4015	0.4983
AR_13	0.001		15.7010	0.0897
AR_14	1	STEF_ICF_method1	15.6820	0.0395
AR_15	10		15.6810	0.0334
AR_16	0.001		15.6825	0.0616
AR_17	1	STEF_ICF_method2	15.6885	0.0572
AR_18	10		15.6715	0.0331
AR_19	0.001		157.4050	3.1162
AR_20	1	STCARlinear	157.6805	1.3098
AR_21	10		157.5290	1.5456
AR_22	0.001		120.0130	0.9894
AR_23	1	STCARanova	120.0150	1.3507
AR_24	10		119.8780	1.0109
AR_25	0.001		646.8420	8.7537
AR_26	1	STCARar	646.9295	2.2494
AR_27	10		648.3700	3.0395

Table 4.19: Percentage of convergent replications of models based on STCAR for scenario AR. Models with Geweke diagnostics values within -2 and 2 are considered as convergent

	$\sigma^2 = 0.001$	$\sigma^2 = 1$	$\sigma^2 = 10$
STCARlinear	0.62	0.66	0.46
STCARanova	0.7	0.68	0.76
STCARar	0.66	0.52	0.72

4.3.4.4 Scenario RF

The numbers of nonzero eigenvectors selected by STEF_CF and STEF_NCF are constantly close to 20 on average, whereas this number reduces to 6 by STEF_ICF (Fig. 4.13). In addition, the number of selected eigenvectors ranges from 0 to 50 for STEF_CF and STEF_NCF whereas it ranges from 0 to 25 from STEF_ICF.

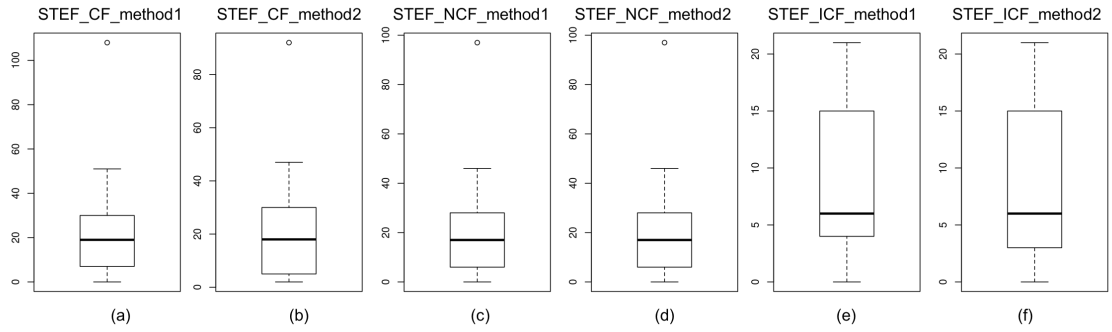

Figure 4.13: Box plots of numbers of nonzero eigenvalues selected for scenario RF by (a) STEF_CF_method1, (b) STEF_CF_method2, (c) STEF_NCF_method1, (d) STEF_NCF_method2, (e) STEF_ICF_method1, and (f) STEF_ICF_method2.

STEF and STCAR parameter estimates for $\hat{\beta}_1$ and $\hat{\beta}_2$ are all close to 1 (Table 4.20). CI widths are close to 0.2 and 0.6 for STEF and STCAR, respectively (Table 4.21). STEF coverage rates range from 56% to 78% whereas the corresponding rates are

Table 4.20: Parameter estimation of Monte Carlo experiment for Scenario RF. Each value is the corresponding median across 50 replicates. Table cells are colored for displaying results based on different modeling methods.

MC experiments	Modeling method	β_1			β_2		
		estimate	95%CI		estimate	95%CI	
RF_1	STEF_CF_method1	1.0215	0.8903	1.1385	0.9740	0.8596	1.1200
RF_2	STEF_CF_method2	1.0119	0.9027	1.1217	0.9811	0.8748	1.0841
RF_3	STEF_NCF_method1	1.0156	0.9136	1.1200	0.9873	0.8831	1.0933
RF_4	STEF_NCF_method2	1.0156	0.9136	1.1200	0.9873	0.8835	1.0933
RF_5	STEF_ICF_method1	1.0184	0.9082	1.1286	0.9783	0.8670	1.0896
RF_6	STEF_ICF_method2	1.0184	0.9082	1.1286	0.9808	0.8707	1.0911
RF_7	STCARlinear	1.0336	0.7097	1.3371	1.0336	0.6422	1.2683
RF_8	STCARanova	1.0384	0.6997	1.3309	0.9515	0.6190	1.2611
RF_9	STCARar	0.9655	0.7083	1.2009	0.9655	0.7850	1.2591

100% for STCAR-derived estimates (Table 4.21). The RMSE and MAE are the lowest using STCARlinear and STCARar, whereas the highest using STEF_CF (Table 4.22). However, the better performance of STCAR comes at a cost. In fact, it may not be preferred given that it takes long time to get results. It is worth mentioning that STEF_NCF and STEF_ICF are not very far from the best performing method while they do not have any convergence issues.

Computational times are close to 28s for STEF_CF and STEF_NCF; STEF_ICF requires about half of that time (Table 4.23). For STCAR, average computational times are equal to 160, 122 and 667 seconds for STCARlinear, STCARanova, and STCARar, respectively. The convergence rates for STCAR models are around 0.6 to 0.7 . Modeling by STCARanova leads to relatively more convergent simulations for Scenario RF (Table 4.24).

Table 4.21: CI length and coverage of Monte Carlo experiment for Scenario RF. Each value is the corresponding median across 50 replicates. Table cells are colored for displaying results based on different methods.

MC experiments	Modeling method	CI length_beta1		CI length_beta2		coverage	
		median	Sd	median	Sd	beta1	beta2
RF_1	STEF_CF_method1	0.3356	0.1079	0.3359	0.1081	56.00%	56.00%
RF_2	STEF_CF_method2	0.2164	0.0080	0.2163	0.0080	72.00%	68.00%
RF_3	STEF_NCF_method1	0.2145	0.0082	0.2145	0.0082	74.00%	78.00%
RF_4	STEF_NCF_method2	0.2145	0.0083	0.2145	0.0083	74.00%	78.00%
RF_5	STEF_ICF_method1	0.2219	0.0226	0.2219	0.0223	76.00%	76.00%
RF_6	STEF_ICF_method2	0.2216	0.0053	0.2216	0.0053	76.00%	78.00%
RF_7	STCARlinear	0.6385	0.1022	0.6278	0.0919	100.00%	91.67%
RF_8	STCARanova	0.6478	0.0779	0.6528	0.0966	100.00%	100.00%
RF_9	STCARar	0.5002	0.0332	0.4963	0.0322	100.00%	87.50%

Table 4.22: RMSE and MAE of Monte Carlo experiment for Scenario RF. Table cells are colored for displaying results based on different methods.

MC experiments	Modeling method	RMSE		MAE	
		β_1	β_2	β_1	β_2
RF_1	STEF_CF_method1	0.2014	0.1074	0.1631	0.0805
RF_2	STEF_CF_method2	0.2068	0.1065	0.1702	0.0855
RF_3	STEF_NCF_method1	0.1008	0.1008	0.0789	0.0789
RF_4	STEF_NCF_method2	0.1000	0.1000	0.0795	0.0795
RF_5	STEF_ICF_method1	0.1117	0.1046	0.0864	0.0825
RF_6	STEF_ICF_method2	0.1047	0.1021	0.0828	0.0809
RF_7	STCARlinear	0.0769	0.0769	0.0631	0.0631
RF_8	STCARanova	0.0863	0.1449	0.0702	0.1120
RF_9	STCARar	0.0842	0.0842	0.0707	0.0707

Table 4.23: Computational time of Monte Carlo experiment for Scenario RF. Table cells are colored for displaying results based on different methods.

MC experiments	Modeling method	computation time	
		median	sd
RF_1	STEF_CF_method1	28.4100	18.4709
RF_2	STEF_CF_method2	28.1415	18.4742
RF_3	STEF_NCF_method1	28.2635	17.4285
RF_4	STEF_NCF_method2	28.0830	17.4515
RF_5	STEF_ICF_method1	14.1995	0.0819
RF_6	STEF_ICF_method2	14.1970	0.0351
RF_7	STCARlinear	160.0530	4.9513
RF_8	STCARanova	122.0965	0.6476
RF_9	STCARar	667.1880	8.9738

Table 4.24: Percentage of convergent replications of models based on STCAR for scenario RF. Models with Geweke diagnostics values within -2 and 2 are considered as convergent.

BMH models	convergence rate
STCARlinear	0.58
STCARanova	0.7
STCARar	0.58

4.3.5 Effects of spatial weight matrix on STEF

To examine the influence of different weight matrices on STEF results, two classical spatial adjacency structures (queen and rook) are compared, focusing on the last Monte Carlo experiment (scenario RF) which does not involve weight matrices in the data generating mechanism (Table 4.25, 4.26, and 4.27). Following the same setting of Monte Carlo experiments in Section 4.3.1., STEF ensemble with rook's adjacency strucure is applied to 50 replicates (hereafter scenario RF_rook). Note that scenario RF (or RF_queen) is exactly based on queen's structure. Therefore scenario RF and scenario RF_rook are compared in this section.

In general, the parameter estimates, confidence intervals, coverages, and precisions are similar across STEFs for different spatial weight matrices. Parameter estimates are close to the true values and CI widths are close to 0.2 in both cases. Coverage rates vary depending on the STEF method but results are similar for different weight matrices. RMSE and MAE are also close, around 0.1 and 0.07, respectively. Therefore, one can conclude that there is no significant influence with respect to spatial weight matrix on the results of STEF.

Table 4.25: Parameter estimation of Monte Carlo experiment for examining effects of spatial weight matrix on STEF. Each value is the corresponding median across 50 replicates. Table cells are colored for displaying results based on different modeling methods.

Spatial weight matrix	Modeling method	β_1			β_2		
		estimate	95% CI		estimate	95% CI	
Queen	STEF_CF_method1	1.0215	0.8903	1.1385	0.9740	0.8596	1.1200
	STEF_CF_method2	1.0119	0.9027	1.1217	0.9811	0.8748	1.0841
	STEF_NCF_method1	1.0156	0.9136	1.1200	0.9873	0.8831	1.0933
	STEF_NCF_method2	1.0156	0.9136	1.1200	0.9873	0.8835	1.0933
	STEF_ICF_method1	1.0184	0.9082	1.1286	0.9783	0.8670	1.0896
	STEF_ICF_method2	1.0184	0.9082	1.1286	0.9808	0.8707	1.0911
Rook	STEF_CF_method1	0.9718	0.8232	1.1263	1.0532	0.8714	1.1865
	STEF_CF_method2	1.0080	0.9017	1.1143	0.9860	0.8781	1.0923
	STEF_NCF_method1	1.0156	0.9099	1.1216	0.9873	0.8777	1.0949
	STEF_NCF_method2	1.0156	0.9099	1.1216	0.9873	0.8777	1.0949
	STEF_ICF_method1	1.0065	0.8976	1.1185	0.9825	0.8720	1.0939
	STEF_ICF_method2	1.0104	0.9009	1.1208	0.9778	0.8660	1.0901

Table 4.26: CI length and coverage of Monte Carlo experiment for examining effects of spatial weight matrix on STEF. Each value is the corresponding median across 50 replicates. Table cells are colored for displaying results based on different methods.

Spatial weight matrix	Modeling method	CI width of β_1		CI width of β_2		coverage	
		median	sd	median	sd	β_1	β_2
Queen	STEF_CF_method1	0.3356	0.1079	0.3359	0.1081	56.00%	56.00%
	STEF_CF_method2	0.2164	0.0080	0.2163	0.0080	72.00%	68.00%
	STEF_NCF_method1	0.2145	0.0082	0.2145	0.0082	74.00%	78.00%
	STEF_NCF_method2	0.2145	0.0083	0.2145	0.0083	74.00%	78.00%
	STEF_ICF_method1	0.2219	0.0226	0.2219	0.0223	76.00%	76.00%
	STEF_ICF_method2	0.2216	0.0053	0.2216	0.0053	76.00%	78.00%
Rook	STEF_CF_method1	0.3659	0.1441	0.3663	0.1411	58.00%	56.00%
	STEF_CF_method2	0.2147	0.0076	0.2158	0.0078	70.00%	68.00%
	STEF_NCF_method1	0.2154	0.0077	0.2154	0.0077	74.00%	78.00%
	STEF_NCF_method2	0.2154	0.0077	0.2154	0.0077	74.00%	78.00%
	STEF_ICF_method1	0.2224	0.0211	0.2235	0.0204	78.00%	74.00%
	STEF_ICF_method2	0.2220	0.0064	0.2228	0.0066	74.00%	76.00%

4.4 Application to bioenergy crop impacts data

The application analyzes an ensemble of large scale spatio-temporal datasets. This ensemble includes WRF-simulated seasonally averaged near-surface temperature differences (°C) over a decade (2000-2009) due to large-scale deployment of perennial bioenergy crops across the continental United States. Two scenarios are included in the ensemble, based on two physics parameterizations under a full deployment

Table 4.27: RMSE and MAE of Monte Carlo experiment for examining effects of spatial weight matrix on STEF. Table cells are colored for displaying results based on different methods.

Spatial weight matrix	Modeling method	RMSE		MAE	
		β_1	β_2	β_1	β_2
Queen	STEF_CF_method1	0.2014	0.1074	0.1631	0.0805
	STEF_CF_method2	0.2068	0.1065	0.1702	0.0855
	STEF_NCF_method1	0.1008	0.1008	0.0789	0.0789
	STEF_NCF_method2	0.1000	0.1000	0.0795	0.0795
	STEF_ICF_method1	0.1117	0.1046	0.0864	0.0825
	STEF_ICF_method2	0.1047	0.1021	0.0828	0.0809
Rook	STEF_CF_method1	0.2357	0.1080	0.1828	0.0800
	STEF_CF_method2	0.2408	0.1094	0.1933	0.0879
	STEF_NCF_method1	0.1008	0.1008	0.0789	0.0789
	STEF_NCF_method2	0.1000	0.1000	0.0795	0.0795
	STEF_ICF_method1	0.1051	0.1076	0.0784	0.0808
	STEF_ICF_method2	0.1137	0.1070	0.0892	0.0850

of perennial bioenergy crops (i.e., E1_100 and E8_100, respectively, as described in Chapter 2). Seasonal averages of this biofuel related datasets over a decade are displayed in Figure 4.14. In this analysis, a specific area located within region 5 (see Figure 4.14) is selected. Region 5 is of interest since it is considered as a sustainable area based on the results presented in Chapter 2. Therefore, the analyzed data (hereafter, T2_biofuel dataset) includes 52 spatial units and 40 temporal units for each scenario, leading to a sample size of 4160. The goal of this application is to quantify the robustness of simulated bioenergy crops impacts (temperature difference) to alternative physics parametrizations.

T2_biofuel is modeled with fixed effects phy_dummy, lat, lon, ele, and seasonal indicator variables (i.e., indicator of using parameter estimation E8 relative to E1, latitude, longitude, elevation, and seasonal indicators, respectively), as well as two-way interactions of lat, lon, and ele. Therefore, 10 fixed effects are included in the model. phy_dummy is of particular interest, as it can be used to quantify the robustness of bioenergy crops impacts. lat, lon, ele, and seasonal indicator variables are included in the model to introduce consistent spatio-temporal information across the two

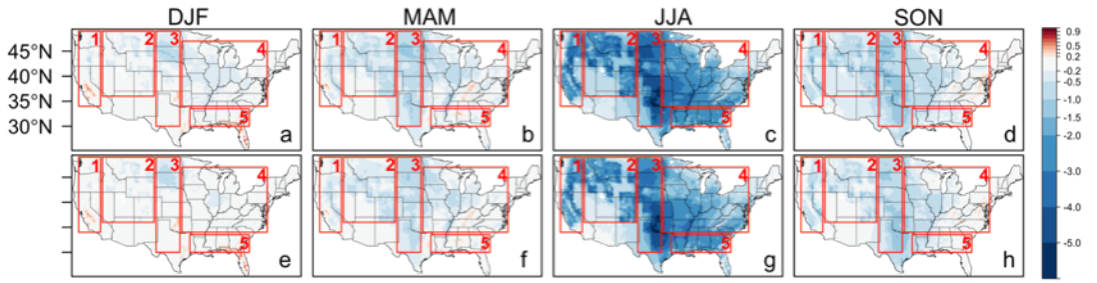

Figure 4.14: WRF-simulated seasonally averaged near-surface temperature difference (°C) over one decade (2000-2009) on (a) DJF, (b) MAM, (c) JJA, and (d) SON using physics parameterization E1. (e)-(h) Same as (a)-(d) but using physics parameterization E8. (Wang et al., 2017).

physics parameterizations. Note that all continuous variables (i.e., lat, lon, ele, and response variable T2_biofuel) are standardized for numerical stability of calculations for parameter estimation. STEF_CF, STEF_NCF, and STEF_ICF approaches are applied to this T2_biofuel dataset. Since results produced by the alternative STEF estimation procedure are similar, method1 is used for this application.

Results of eigenvector selection and computational times differ across STEFs. The number of selected eigenvectors are 469s and 401s for STEF_CF and STEF_NCF, respectively, whereas only 47 for STEF_ICF (Table 4.28). Nevertheless, absolute magnitudes of estimated eigenvectors coefficients are consistently small (around 0.2), compared to around 1 to 2 for fixed effects (Figure 4.15). Regarding to computational times, STEF_CF and STEF_NCF procedures are close, by around 293.6s and 281.5s, respectively. However, STEF_ICF only requires 60.6s (Table 4.28).

Parameter estimates and significance of fixed effects are partially consistent across STEFs (Table 4.28). Variables phy_dummy, lon*ele, Spring-Winter, Summer-Winter, and Fall-Winter are statistically significant. This result is confirmed by STEF_CF,

Table 4.28: Parameter estimates, 95% confidence intervals, number of selected eigenvectors and computational times for modeling T2_biofuel dataset by STEF. Statistically significant variables are highlighted in red.

	STEF_CF			STEF_NCF			STEF_ICF		
	Estimate	95% CI		Estimate	95% CI		Estimate	95% CI	
phy_dummy	0.6296	0.5955	0.6636	0.3370	0.3313	0.3428	0.1341	0.1168	0.1514
lon	0.3730	-0.0349	0.7809	-0.1783	-0.3972	0.0407	0.0007	-0.0084	0.0099
lat	-0.9286	-1.8822	0.0250	0.2836	-0.2292	0.7965	-0.0036	-0.0186	0.0115
ele	1.3060	0.2185	2.3935	1.7740	1.1438	2.4042			
lon*lat	-0.9143	-1.9210	0.0924	0.3130	-0.2323	0.8583			
lon*ele	1.1695	0.2434	2.0956	1.2911	0.7503	1.8318	0.0530	0.0374	0.0687
lat*ele	-0.1223	-0.4424	0.1978	-0.5079	-0.6555	-0.3603			
Spring-Winter	0.3038	0.2723	0.3352	0.6325	0.6254	0.6396	0.5330	0.5017	0.5644
Summer-Winter	-1.6329	-1.6665	-1.5992	-1.6668	-1.6739	-1.6597	-1.4194	-1.4475	-1.3912
Fall-Winter	0.0608	0.0365	0.0850	0.3459	0.3387	0.3530	0.4102	0.3779	0.4426
Number of nonzero eigenvectors	469			401			47		
Computational time	293.6			281.541			60.5800		

STEF_NCF, and STEF_ICF as these significant parameter estimates possess the same sign (positive or negative) for all STEFs. Besides this agreement, STEF_ICF excludes variables ele, lon*lat, lat*ele due to their colinearity with other fixed effects. For those STEF_ICF-excluded fixed effects, variable lat*ele is not statistically significant by STEF_CF whereas it is statistically significant by STEF_NCF; parameter estimate of lon*lat by STEF_CF is negative in contrast with STEF_NCF. These results suggest that outputs from STEF_CF are affected by multicollinearity. The confidence interval lengths of fixed effects are shorter for STEF_NCF and STEF_ICF (except for dummy variable Fall-Winter, see Table 4.29), which most probably is also due to multicollinearity. When comparing the differences of parameter estimates (i.e., deviance), STEF_ICF provides estimates which are close to STEF_NCF for phy_dummy, lon, lat, Spring-Winter, and Fall-Winter (note that most of them are significant variables).

STEF_ICF alleviates the spatial confounding effects and achieves high accuracy without endorsing the assumptions of STEF_NCF. The effects of different physics parameterizations are statistically significant although, not of the same practical

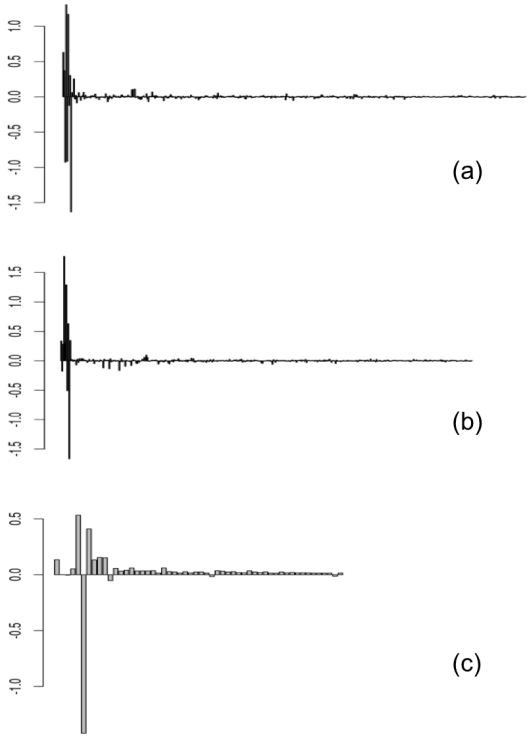

Figure 4.15: Barplot of parameter estimates for modeling T2_biofuel dataset by (a) STEF_CF, (b) STEF_NCF, and (c) STEF_ICF, respectively. For each panel, the first 10 bars on the left, and the bars beginning from the 11^{th} bars to the right represent the parameter estimates for fixed effects, and eigenvectors, respectively.

significance compared to the effects of environmental and geographical predictors. Controlling the other variables, temperature change associated with perennial bioenergy crops is 0.13°C higher using parameterization E8 than using parameterization E1.[2] Spatially, increasing one unit standard deviation of longitude and one unit standard deviation of elevation leads to temperature increasing by 0.052°C. Comparing

[2]Since continuous explanatory variables and response variables are all standardized, the estimated coefficients are scaled back to get the unit changes in original units of data. For example, 0.13 is calculated by multiplying parameter estimate phy_dummy (0.13) to the standard deviation of response variable (0.98).

Table 4.29: CI width and deviance of parameter estimates for modeling T2_biofuel dataset by STEF. The deviances of CF-NCF and ICF-NCF are calculated as the parameter estimate differences between using STEF_CF and STEF_NCF, and between using STEF_ICF and STEF_NCF, respectively. Smallest and median CI widths among STEF specifications are highlighted in red and blue, respectively; smaller deviance among STEF specifications are highlighted in red; variables in gray are correlated variables identified by STEF_ICF, which do not included in the analysis.

	CI width			Deviance	
	STEF_CF	STEF_NCF	STEF_ICF	CF-NCF	ICF-NCF
phy_dummy	0.0680	0.0115	0.0346	0.2925	-0.2030
lon	0.8158	0.4379	0.0183	0.5513	0.1790
lat	1.9072	1.0257	0.0301	-1.2122	-0.2872
ele	2.1750	1.2604		-0.4680	
lon*lat	2.0134	1.0906		-1.2273	
lon*ele	1.8522	1.0815	0.0313	-0.1216	-1.2380
lat*ele	0.6402	0.2952		0.3856	
Spring-Winter	0.0629	0.0143	0.0627	-0.3287	-0.0994
Summer-Winter	0.0674	0.0143	0.0563	0.0340	0.2475
Fall-Winter	0.0485	0.0143	0.0646	-0.2851	0.0644

to Winter, cooling impacts associated with perennial bioenergy crop expansion are 0.52°C and 0.4°C lower in Spring and Fall, respectively; whereas 1.39°C higher in Fall.

4.5 Discussion

This Chapter developed a framework for modeling space-time lattice data. Three approaches of STEF- introducing proxy variables with, without, or with intermediate spatial confounding - are evaluated. In addition, two alternative algorithms for implementing each STEF method, are explored. LSA to Adaptive Lasso for eigenvector selection can be considered as a method for consistent parameter estimation, which increases computational efficiency. Applying VIF filtering and SIS screening reduce the number of correlated explanatory variables and eigenvectors in the model, resulting in a parsimonious specification. Most importantly, STEF with VIF-SIS approach conducts similar results of BHM (i.e., STCARlinear, STCARanova, and STCARar)

yet is much more computationally efficient. Therefore, STEF can be used as a more reliable method for modeling spatio-temporal data. However, STEF_CF_method1 does not perform well under variety of spatio-temporal autocorrelation structure, so other STEF methods should be considered. Seasons are statistically significant factors for temperature changes associated with perennial bioenergy crops over the east coast of the US. Thus our models suggest that the impacts of bioenergy crops are neither fixed in time nor fixed in space. This region shows statistically significantly different temperature impact, modeled by WRF with alternative physics parameterizations.

The spatial confounding effect is verified at different scales and at different spatio-temporal structures in this study. For scenario AR or RF which do not in favor of STEF algorithm, RMSE and MAE of STEF_CF are higher while CI widths of STEF_CF are larger, relative to STEF_NCF. Therefore, taking into account spatial confounding (i.e., using STEF_NCF rather than STEF_CF) could not only increase the accuracy of the estimation, but also reduce the variance inflation of parameter estimates. STEF_ICF, on the other hand, eliminated some eigenvectors that are strongly correlated with the fixed-effects but does not fully eliminate spatial confounding. Hence STEF_ICF lies between STEF_CF and STEF_NCF with regard to estimation accuracy and variance as measured by RMSE, MAE, and CI width.

The coverage rate varies across modeling method and spatio-temporal correlation structure of data. This uncertainty of coverage rate may due to the combined effect of bias and variance. One would expect better coverage for smalle values of σ^2. However, this is not what we observe consistently for STEF methods. Considering the spatial confounding scenario, the variance inflated leads to CI width increases, resulting the possibility of covering larger of range of values. On the other hand, eigenvectors that correlated with fixed effects variables may be introduced in the

model, causing the estimation of fixed effects to be biased. If this bias is so large that the wider confidence interval cannot compensate, the coverage would still be low. When modeling with STEF_ICF, the extra step of alleviating spatial confounding can be considered as a third effect, thus the relative order of capability of coveraging true values among STEF_CF, STEF_NCF, and STEF_ICF differs for data with different spatio-temporal correlation. For STCAR models, coverage do not appear to depend on σ^2 neither ; most probably this is due to small sample sizes as several MCMC procedures do not converge.

STEF show advantage of no assumption of underlying spatio-temporal structure and relatively simple implementation, resulting in robust parameter estimates with high computational efficiency. STEF performs well when the spatio-temporal autocorrelation structure is in favor of eivenvector filtering algorithm. The well performance for EFM1 and EFMX scenario verifies that STEF is capable of capturing latent spatio-temporal autocorrelation structure at a variety of scale, resulting more precise estimation and coverage, especially when spatial confounding is taken into account. With the ability of capturing latent proxy structures, STEF fits a more general spatial temporal model, which do not depends on the assumption of covariance structure of the data. This property prevents model to be fitted too restricted to capture the true value. Based on our study, STEF models (especially STEF_NCF and STEF_ICF) still perform close to STCAR models even if the spatio-temporal strucure is more complicated . The computational time of STEF, however, is around one tenth of STCAR. It is worth noting that STCAR models only converge in a roughly of half of modeling over long MCMC samplings, which reduce the valid replicates in Monte Carlo experiments. Therefore, STEF has advantage of robust estimation with low computational burden.

The computational burden increases when one moves from spatial to spatio-temporal analyses. Computational times for eigenvector filtering depend largely on the speed of eigendecomposition for large datasets. Some eigendecomposition techniques have been utilized, such as a approximate eigendecompostion based on Nystroom extension by Murakami and Griffith (2017). In this Chapter, eigenvector decomposition for sparse matrix is using R funtion RSpectra is suggested, aiming to improve the capability of STEF for analyzing larger spatio-temporal data. Computational efficiency is also determined by Adaptive Lasso, which computational time could increase as more variables are included in the model. Using VIF-SIS procedures within STEF_ICF, the model is more parsimonious by excluding correlated fixed effects and eigenvectors, thus the computational time is largely reduced.

The application result using STEF differ from the results obtained using BHM models in Chapter 3. In particularly, the physcis parameterization are statistically significant based on STEF, whereas not significant using BHM. Fundamentally, these two results are not comparable as the data in different resolutions and for different regions. For BHM modeling, a resolution of 2.5 °C over the Continental US are used, whereas 0.5 degeree in local area (within Region 5) is of focus. Nevertheless, given the superior in capturing spatio-temporal autocorrelation structure of data with high efficiency, STEF could provide a more reliable result. It could be possible that STEF show statistically nonsignificance of effect of physics parameterization over the entire U.S., although the locally difference has been found. In fact, this discrepancy of results indicate the need of studying by spatial varying coefficient to capture spatio-temporal structure in local.

It is worth noting that using projection matrix $\boldsymbol{M}_{(X)}$ to generate eigenvectors is under the assumption that all omitted (weak) predictors are orthogonal to the

predictors in $\boldsymbol{X}$. Spatial correlation may arise due to omitted (weak) predictors which are spatially correlated: this is indeed the spatial error model. The eigenvectors included in STEF could be considered to explain the effects of omitted predictors. However, it is possible that there exists some extra predictors should be included in the model and they are correlated with the already included predictors. In this case, the assumption of eigenvectors based on $\boldsymbol{M}_{(X)}$ as omitted predictors are not valid as these eigenvectors are orthogonal to the predictors. Therefore, STEF_ICF first alleviates the restriction of perpendicular of predictors and eigenvectors, then reduce the collinearity could seems as a more general way. In fact, the Monte Carlo experiments and application all show the high accuracy and coverage of STEF_ICF without restricted assumption of orthogonality.

Some limitations of STEF should be mentioned. For instance, the results of STEF depend on the criteria and methods of eigenvector selection. Changing the criteria to determine the multicolinearity or regularization approaches may cause a different specification to capture spatio-temporal correlation, leading to a possibly different estimates. In addition, the conclusion about computational efficiency does not take into account the sample size of data. A thorough analysis on different scale of sample size is needed. Lastly, this method can only be study global parameter. In this case one would have to consider extended versions of STEF, which include interaction terms based on fixed effects and Moran eigenvectors.

Chapter 5

CONCLUSIONS AND DIRECTIONS FOR FURTHER WORK

The hydroclimatic impacts associated with perennial bioenergy crop expansion over the contiguous United States, are quantified using the Weather Research and Forecasting Model, dynamically coupled to a land surface model (LSM). A suite of continuous (2000–09) medium-range resolution (20-km grid spacing) ensemble-based simulations is conducted using a seasonally evolving biophysical representation of perennial bioenergy cropping systems within the LSM based on observational data. Deployment is carried out only over suitable abandoned and degraded farmlands to avoid competition with existing food cropping systems.

The WRF simulation results show that near-surface cooling (locally, up to 58 °C) is greatest during the growing season over portions of the central United States. For some regions, principal impacts are restricted to a reduction in near-surface temperature (e.g., eastern portions of the United States), whereas for other regions deployment leads to soil moisture reduction in excess of 0.15–0.2 m^3 m^{-3} during the simulated 10-yr period (e.g., western Great Plains). This reduction (25%–30% of available soil moisture) manifests as a progressively decreasing trend over time. The large-scale focus of this research demonstrates the long-term hydroclimatic sustainability of large-scale deployment of perennial bioenergy crops across the continental United States, revealing potential hot spots of suitable deployment and regions to avoid.

Hovmöller and Taylor diagrams are utilized to evaluate simulated temperature and precipitation. Using this technique, a quantitative analysis of model performance is conducted. The best and least skilled physics parametrizations are selected for

bioenergy crop expansion simulations. In addition, Mann–Kendall modified trend tests and Sieve-bootstrap trend tests are used to evaluate the statistical significance of trends in soil moisture differences. These two types of tests show consistent results of statistically significant decreasing trends in soil moisture. Based on the aforementioned analysis, potential hot spots of suitable deployment and regions to avoid are determined.

Besides explanatory analysis of model performance, the robustness of WRF simulations to alternative physics parametrizations is evaluated using Bayesian Hierarchical spatio-temporal models. Specifications with spatially varying intercepts and slopes can offer a satisfactory description of the spatio-temporal dependence structure of the data. Simulated impacts on temperatures due to perennial bioenergy crop expansion are found robust to physics parameterizations: the main findings of the analysis do not change significantly with alternative parametrizations.

Given the computational burden of BHM, a spatio-temporal eigenvectoring filtering scheme is proposed as a computationally efficient modeling approach. Three conditions - introducing proxy variables with, without, or with intermediate spatial confounding - are explored. In addition, two approaches for two-step STEF are utilized. VIF-based filtering and Sure Independence Screening can reduce the total number of eigenvectors in the model selection procedure, resulting in more accurate estimates. The least squares approximation to Adaptive Lasso for eigenvector selection has been used to obtain significant eigenvectors, aiming at consistent parameter estimation and computational efficiency. STEF has shown superior for data in accordance with the STEF data generating mechanism. For modeling data with other spatio-temporal autocorrelation structure, STEF_NCF and STEF_ICF are still being suggested, as they are not very far from the best performing method. At the same time they

do not have convergence issues and they do not take long time to converge. More importantly, STEF_ICF alleviates the spatial confounding effects and achieves high accuracy without endorsing the assumptions of STEF_NCF.

Future studies may extend the work presented in this thesis. For example, STEF method proposed in this thesis can be easily extended to take into account spatially varying coefficients, by including interactions between fixed effects and the constructed Moran eigenvectors. However, these interactions may be colinear with the fixed effects in the model; STEF is expected to be valuable in this situation, as problematic eigenvectors can be removed through filtering and screening and eigenvector selection can be more accurate and fast. This model selection procedure proposed in this study are useful to coupled with other statistical method, i.e., quantile regression, to increase the precision. In addition, the comparison of geographically weighted regression (McMillen 2004) with STEF for spatio-temporal datasets should be examined. These topics are left for future research.

With regard to the application in the analysis of outputs from regional climate models, one is interested in evaluating whether some parametrizations dominate in regions of the examined spatial domain and whether ensemble schemes may be created based on such properties. STEF with spatially varying coefficients can be used to address such research questions. In addition, STEF modeling on different scale of regional climate models outputs should be conducted to examine different scales of phenomenon. These topics remains to be addressed in future research efforts.

NOTES

REFERENCES

Abraha, M., J. Chen, H. Chu, T. Zenone, R. John, Y.-J. Su, S. K. Hamilton and G. P. Robertson, "Evapotranspiration of annual and perennial biofuel crops in a variable climate", Gcb Bioenergy **7**, 6, 1344–1356 (2015).

Akaike, H., "Information theory and an extension of the maximum likelihood principle", in "Selected Papers of Hirotugu Akaike", pp. 199–213 (Springer, 1998).

Anderson, C. J., R. P. Anex, R. W. Arritt, B. K. Gelder, S. Khanal, D. E. Herzmann and P. W. Gassman, "Regional climate impacts of a biofuels policy projection", Geophysical Research Letters **40**, 6, 1217–1222 (2013).

Anderson-Teixeira, K. J., S. C. Davis, M. D. Masters and E. H. Delucia, "Changes in soil organic carbon under biofuel crops", Gcb Bioenergy **1**, 1, 75–96 (2009).

Anderson-Teixeira, K. J., P. K. Snyder, T. E. Twine, S. V. Cuadra, M. H. Costa and E. H. DeLucia, "Climate-regulation services of natural and agricultural ecoregions of the americas", Nature Climate Change **2**, 3, 177 (2012).

Anselin, L. and D. A. Griffith, "Do spatial effecfs really matter in regression analysis?", Papers in Regional Science **65**, 1, 11–34 (1988).

Anselin, L. and S. Rey, "Properties of tests for spatial dependence in linear regression models", Geographical analysis **23**, 2, 112–131 (1991).

Aragon, N. U., M. Wagner, M. Wang, A. M. Broadbent, N. Parker and M. Georgescu, "Sustainable land management for bioenergy crops", Energy Procedia **125**, 379–388 (2017).

Bagley, J. E., S. C. Davis, M. Georgescu, M. Z. Hussain, J. Miller, S. W. Nesbitt, A. VanLoocke and C. J. Bernacchi, "The biophysical link between climate, water, and vegetation in bioenergy agro-ecosystems", biomass and bioenergy **71**, 187–201 (2014).

Bakar, K. S., P. Kokic and H. Jin, "Hierarchical spatially varying coefficient and temporal dynamic process models using sptdyn", Journal of Statistical Computation and Simulation **86**, 4, 820–840 (2016).

Barut, E., J. Fan and A. Verhasselt, "Conditional sure independence screening", Journal of the American Statistical Association **111**, 515, 1266–1277 (2016).

Bates, D., M. Mächler, B. Bolker and S. Walker, "Fitting linear mixed-effects models using lme4", arXiv preprint arXiv:1406.5823 (2014).

Belloni, A., V. Chernozhukov *et al.*, "Least squares after model selection in high-dimensional sparse models", Bernoulli **19**, 2, 521–547 (2013).

Bernardinelli, L., D. Clayton, C. Pascutto, C. Montomoli, M. Ghislandi and M. Songini, "Bayesian analysis of space?time variation in disease risk", Statistics in medicine **14**, 21-22, 2433–2443 (1995).

Besag, J., J. York and A. Mollie, "Bayesian image restoration with two applications in spatial statistics (with discussion) ann inst stat math. 1991; 43: 1–59. doi: 10.1007", BF00116466.[Cross Ref] (1991).

Betts, R. A., "Offset of the potential carbon sink from boreal forestation by decreases in surface albedo", Nature **408**, 6809, 187 (2000).

Blangiardo, M. and M. Cameletti, *Spatial and spatio-temporal Bayesian models with R-INLA* (John Wiley & Sons, 2015).

Cai, X., X. Zhang and D. Wang, "Land availability for biofuel production", Environmental science & technology **45**, 1, 334–339 (2010).

Caiazzo, F., R. Malina, M. D. Staples, P. J. Wolfe, S. H. Yim and S. R. Barrett, "Quantifying the climate impacts of albedo changes due to biofuel production: a comparison with biogeochemical effects", Environmental Research Letters **9**, 2, 024015 (2014).

Campbell, J. E., D. B. Lobell, R. C. Genova and C. B. Field, "The global potential of bioenergy on abandoned agriculture lands", Environmental science & technology **42**, 15, 5791–5794 (2008).

Campbell, J. E., D. B. Lobell, R. C. Genova, A. Zumkehr and C. B. Field, "Seasonal energy storage using bioenergy production from abandoned croplands", Environmental Research Letters **8**, 3, 035012 (2013).

Chen, C.-T. and T. Knutson, "On the verification and comparison of extreme rainfall indices from climate models", Journal of Climate **21**, 7, 1605–1621 (2008).

Chen, F. and J. Dudhia, "Coupling an advanced land surface–hydrology model with the penn state–ncar mm5 modeling system. part i: Model implementation and sensitivity", Monthly Weather Review **129**, 4, 569–585 (2001).

Christian, P. R. and G. Casella, "Monte carlo statistical methods", (1999).

Chun, Y., "Modeling network autocorrelation within migration flows by eigenvector spatial filtering", Journal of Geographical Systems **10**, 4, 317–344 (2008).

Chun, Y., "Analyzing space–time crime incidents using eigenvector spatial filtering: an application to vehicle burglary", Geographical Analysis **46**, 2, 165–184 (2014).

Chun, Y. and D. A. Griffith, "Modeling network autocorrelation in space–time migration flow data: an eigenvector spatial filtering approach", Annals of the Association of American Geographers **101**, 3, 523–536 (2011).

Chun, Y. and D. A. Griffith, *Spatial statistics and geostatistics: theory and applications for geographic information science and technology* (Sage, 2013).

Chun, Y. and D. A. Griffith, "A quality assessment of eigenvector spatial filtering based parameter estimates for the normal probability model", Spatial Statistics **10**, 1–11 (2014).

Cliff, A. D. and J. K. Ord, "Spatial and temporal analysis: autocorrelation in space and time", Quantitative geography: a British view pp. 104–110 (1981).

Clifton-Brown, J. C., J. Breuer and M. B. Jones, "Carbon mitigation by the energy crop, miscanthus", Global Change Biology **13**, 11, 2296–2307 (2007).

Council, N. R. *et al.*, *Liquid transportation fuels from coal and biomass: technological status, costs, and environmental impacts* (National Academies Press, 2010).

Cox, D. R. and A. Stuart, "Some quick sign tests for trend in location and dispersion", Biometrika **42**, 1/2, 80–95 (1955).

Cressie, N. and C. K. Wikle, *Statistics for spatio-temporal data* (John Wiley & Sons, 2015).

Davin, E. L., S. I. Seneviratne, P. Ciais, A. Olioso and T. Wang, "Preferential cooling of hot extremes from cropland albedo management", Proceedings of the National Academy of Sciences **111**, 27, 9757–9761 (2014).

DeLucia, E. H., "How biofuels can cool our climate and strengthen our ecosystems", Eos **96**, 4, 14–19 (2015).

Dohleman, F. G. and S. P. Long, "More productive than maize in the midwest: how does miscanthus do it?", Plant physiology **150**, 4, 2104–2115 (2009).

Dondini, M., K.-J. Van Groenigen3, I. Del Galdo and M. B. Jones, "Carbon sequestration under miscanthus: a study of 13c distribution in soil aggregates", Gcb Bioenergy **1**, 5, 321–330 (2009).

Done, J. M., L. R. Leung, C. A. Davis and B. Kuo, "Simulation of warm season rainfall using wrf regional climate model", in "6th WRF/15th MM5 users? workshop, Boulder, CO, USA", (2005).

Efron, B., T. Hastie, I. Johnstone, R. Tibshirani *et al.*, "Least angle regression", The Annals of statistics **32**, 2, 407–499 (2004).

Eichelmann, E., C. Wagner-Riddle, J. Warland, B. Deen and P. Voroney, "Comparison of carbon budget, evapotranspiration, and albedo effect between the biofuel crops switchgrass and corn", Agriculture, Ecosystems & Environment **231**, 271–282 (2016).

Ek, M., K. Mitchell, Y. Lin, E. Rogers, P. Grunmann, V. Koren, G. Gayno and J. Tarpley, "Implementation of noah land surface model advances in the national centers for environmental prediction operational mesoscale eta model", Journal of Geophysical Research: Atmospheres **108**, D22 (2003).

Fan, J. and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties", Journal of the American statistical Association **96**, 456, 1348–1360 (2001).

Fan, J. and J. Lv, "Sure independence screening for ultrahigh dimensional feature space", Journal of the Royal Statistical Society: Series B (Statistical Methodology) **70**, 5, 849–911 (2008).

Fan, Y. and H. Van den Dool, "A global monthly land surface air temperature analysis for 1948–present", Journal of Geophysical Research: Atmospheres **113**, D1 (2008).

Fargione, J., J. Hill, D. Tilman, S. Polasky and P. Hawthorne, "Land clearing and the biofuel carbon debt", Science **319**, 5867, 1235–1238 (2008).

Feng, Q., I. Chaubey, Y. G. Her, R. Cibin, B. Engel, J. Volenec and X. Wang, "Hydrologic and water quality impacts and biomass production potential on marginal land", Environmental Modelling & Software **72**, 230–238 (2015).

Ferchaud, F., G. Vitte, F. Bornet, L. Strullu and B. Mary, "Soil water uptake and root distribution of different perennial and annual bioenergy crops", Plant and soil **388**, 1-2, 307–322 (2015).

Field, C. B., J. E. Campbell and D. B. Lobell, "Biomass energy: the scale of the potential resource", Trends in ecology & evolution **23**, 2, 65–72 (2008).

Finley, A. O., S. Banerjee and A. E. Gelfand, "spbayes for large univariate and multivariate point-referenced spatio-temporal data models", arXiv preprint arXiv:1310.8192 (2013).

Fomby, T. B. and T. J. Vogelsang, "The application of size-robust trend statistics to global-warming temperature series", Journal of Climate **15**, 1, 117–123 (2002).

Gelfand, I., R. Sahajpal, X. Zhang, R. C. Izaurralde, K. L. Gross and G. P. Robertson, "Sustainable bioenergy production from marginal lands in the us midwest", Nature **493**, 7433, 514 (2013).

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari and D. B. Rubin, *Bayesian data analysis*, vol. 2 (CRC press Boca Raton, FL, 2014).

Geman, S. and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images", in "Readings in Computer Vision", pp. 564–584 (Elsevier, 1987).

Georgescu, M., D. Lobell and C. Field, "Potential impact of us biofuels on regional climate", Geophysical Research Letters **36**, 21 (2009).

Georgescu, M., D. Lobell, C. Field and A. Mahalov, "Simulated hydroclimatic impacts of projected brazilian sugarcane expansion", Geophysical Research Letters **40**, 5, 972–977 (2013).

Georgescu, M., D. B. Lobell and C. B. Field, "Direct climate effects of perennial bioenergy crops in the united states", Proceedings of the National Academy of Sciences **108**, 11, 4307–4312 (2011).

Getis, A. and D. A. Griffith, "Comparative spatial filtering in regression analysis", Geographical analysis **34**, 2, 130–140 (2002).

Geweke, J. *et al.*, *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*, vol. 196 (Federal Reserve Bank of Minneapolis, Research Department Minneapolis, MN, USA, 1991).

Gneiting, T., "Nonseparable, stationary covariance functions for space–time data", Journal of the American Statistical Association **97**, 458, 590–600 (2002).

Goldstein, J. C., A. Tarhule and D. Brauer, "Simulating the hydrologic response of a semiarid watershed to switchgrass cultivation", Hydrology Research **45**, 1, 99–114 (2014).

Grell, G. A., "Prognostic evaluation of assumptions used by cumulus parameterizations", Monthly Weather Review **121**, 3, 764–787 (1993).

Grell, G. A. and D. Dévényi, "A generalized approach to parameterizing convection combining ensemble and data assimilation techniques", Geophysical Research Letters **29**, 14 (2002).

Griffith, D., "Interdependence in space and time: numerical and interpretative considerations", Dynamic spatial models pp. 258–287 (1981).

Griffith, D., "Spatial filtering and missing georeferenced data imputation: A comparison of the getis and griffith methods", in "Perspectives on spatial data analysis", pp. 227–233 (Springer, 2010).

Griffith, D., "Space, time, and space-time eigenvector filter specifications that account for autocorrelation", Estadística española **54**, 177, 7–34 (2012).

Griffith, D. A., "Spatial autocorrelation and eigenfunctions of the geographic weights matrix accompanying geo-referenced data", The Canadian Geographer/Le Géographe canadien **40**, 4, 351–367 (1996).

Griffith, D. A., "A linear regression solution to the spatial autocorrelation problem", Journal of Geographical Systems **2**, 2, 141–156 (2000).

Griffith, D. A., "A spatial filtering specification for the auto-poisson model", Statistics & probability letters **58**, 3, 245–251 (2002).

Griffith, D. A., *Spatial autocorrelation and spatial filtering: gaining understanding through theory and scientific visualization* (Springer Science & Business Media, 2003).

Griffith, D. A., "Distributional properties of georeferenced random variables based on the eigenfunction spatial filter", Journal of geographical systems **6**, 3, 263–288 (2004a).

Griffith, D. A., "A spatial filtering specification for the autologistic model", Environment and Planning A **36**, 10, 1791–1811 (2004b).

Griffith, D. A., "Effective geographic sample size in the presence of spatial autocorrelation", Annals of the Association of American Geographers **95**, 4, 740–760 (2005).

Griffith, D. A., "Visualizing analytical spatial autocorrelation components latent in spatial interaction data: an eigenvector spatial filter approach", Computers, Environment and Urban Systems **35**, 2, 140–149 (2011).

Griffith, D. A., "Some robustness assessments of moran eigenvector spatial filtering", Spatial Statistics **22**, 155–179 (2017).

Griffith, D. A. and Y. Chun, "Spatial analysis of census mail response rates: 1990–2010", in "Space-Time Integration in Geography and GIScience", pp. 145–156 (Springer, 2015).

Griffith, D. A. and M. M. Fischer, "Constrained variants of the gravity model and spatial dependence: model specification and estimation issues", in "Spatial Econometric Interaction Modelling", pp. 37–66 (Springer, 2016).

Griffith, D. A. and J. H. Paelinck, *Morphisms for Quantitative Spatial Analysis*, vol. 51 (Springer, 2018).

Griffith, D. A. and J. H. P. Paelinck, *Non-standard spatial statistics and spatial econometrics* (Springer Science & Business Media, 2011).

Griffith, D. A. and P. R. Peres-Neto, "Spatial modeling in ecology: the flexibility of eigenfunction spatial analyses", Ecology **87**, 10, 2603–2613 (2006).

Hallgren, W., C. A. Schlosser, E. Monier, D. Kicklighter, A. Sokolov and J. Melillo, "Climate impacts of a large-scale biofuels expansion", Geophysical Research Letters **40**, 8, 1624–1630 (2013).

Hamed, K. H. and A. R. Rao, "A modified mann-kendall trend test for autocorrelated data", Journal of Hydrology **204**, 1-4, 182–196 (1998).

Hanks, E. M., E. M. Schliep, M. B. Hooten and J. A. Hoeting, "Restricted spatial regression in practice: geostatistical models, confounding, and robustness under model misspecification", Environmetrics **26**, 4, 243–254 (2015).

Hastings, W. K., "Monte carlo sampling methods using markov chains and their applications", (1970).

Hefley, T. J., M. B. Hooten, E. M. Hanks, R. E. Russell and D. P. Walsh, "The bayesian group lasso for confounded spatial data", Journal of Agricultural, Biological and Environmental Statistics **22**, 1, 42–59 (2017).

Helbich, M. and D. A. Griffith, "Spatially varying coefficient models in real estate: Eigenvector spatial filtering and alternative approaches", Computers, Environment and Urban Systems **57**, 1–11 (2016).

Hickman, G. C., A. Vanloocke, F. G. Dohleman and C. J. Bernacchi, "A comparison of canopy evapotranspiration for maize and two perennial grasses identified as potential bioenergy crops", Gcb Bioenergy **2**, 4, 157–168 (2010).

Higgins, R. W., W. Shi, E. Yarosh and R. Joyce, "Improved united states precipitation quality control system and analysis", NCEP/Climate prediction center atlas **7**, 40 (2000).

Hodges, J. S. and B. J. Reich, "Adding spatially-correlated errors can mess up the fixed effect you love", The American Statistician **64**, 4, 325–334 (2010).

Hoeting, J. A., D. Madigan, A. E. Raftery and C. T. Volinsky, "Bayesian model averaging: a tutorial", Statistical science pp. 382–401 (1999).

Hong, S.-Y., J. Dudhia and S.-H. Chen, "A revised approach to ice microphysical processes for the bulk parameterization of clouds and precipitation", Monthly Weather Review **132**, 1, 103–120 (2004).

Hovmöller, E., "The trough-and-ridge diagram", Tellus **1**, 2, 62–66 (1949).

Hudiburg, T. W., S. C. Davis, W. Parton and E. H. Delucia, "Bioenergy crop greenhouse gas mitigation potential under a range of management practices", Gcb Bioenergy **7**, 2, 366–374 (2015).

Hudiburg, T. W., W. Wang, M. Khanna, S. P. Long, P. Dwivedi, W. J. Parton, M. Hartman and E. H. DeLucia, "Impacts of a 32-billion-gallon bioenergy landscape on land and fossil fuel use in the us", Nature Energy **1**, 1, 15005 (2016).

Hughes, J. and M. Haran, "Dimension reduction and alleviation of confounding for spatial generalized linear mixed models", Journal of the Royal Statistical Society: Series B (Statistical Methodology) **75**, 1, 139–159 (2013).

Kain, J. S., "The kain–fritsch convective parameterization: an update", Journal of applied meteorology **43**, 1, 170–181 (2004).

Kang, E. L., N. Cressie and S. R. Sain, "Combining outputs from the north american regional climate change assessment program by using a bayesian hierarchical model", Journal of the Royal Statistical Society: Series C (Applied Statistics) **61**, 2, 291–313 (2012).

Kendall, M. G., "Rank correlation methods", (1955).

Khaliq, M., T. B. Ouarda, P. Gachon, L. Sushama and A. St-Hilaire, "Identification of hydrological trends in the presence of serial and cross correlations: A review of selected methods and their application to annual flow regimes of canadian rivers", Journal of Hydrology **368**, 1-4, 117–130 (2009).

Khanal, S., R. P. Anex, C. J. Anderson and D. E. Herzmann, "Streamflow impacts of biofuel policy-driven landscape change", PloS one **9**, 10, e109129 (2014).

Khanal, S., R. P. Anex, C. J. Anderson, D. E. Herzmann and M. K. Jha, "Implications of biofuel policy-driven land cover change for rainfall erosivity and soil erosion in the united states", Gcb Bioenergy **5**, 6, 713–722 (2013).

Kisi, O. and M. Ay, "Comparison of mann–kendall and innovative trend method for water quality parameters of the kizilirmak river, turkey", Journal of Hydrology **513**, 362–375 (2014).

Knorr-Held, L., "Bayesian modelling of inseparable space-time variation in disease risk", (1999).

Knorr-Held, L., "Dynamic rating of sports teams", Journal of the Royal Statistical Society: Series D (The Statistician) **49**, 2, 261–276 (2000).

Knorr-Held, L. and J. Besag, "Modelling risk from a disease in time and space", Statistics in medicine **17**, 18, 2045–2060 (1998).

Kundzewicz, Z. W. and A. J. Robson, "Change detection in hydrological records?a review of the methodology/revue méthodologique de la détection de changements dans les chroniques hydrologiques", Hydrological sciences journal **49**, 1, 7–19 (2004).

Le, P. V., P. Kumar and D. T. Drewry, "Implications for the hydrologic cycle under climate change due to the expansion of bioenergy crops in the midwestern united states", Proceedings of the National Academy of Sciences **108**, 37, 15085–15090 (2011).

Lee, D. and A. Lawson, "Quantifying the spatial inequality and temporal trends in maternal smoking rates in glasgow", The annals of applied statistics **10**, 3, 1427 (2016).

Lee, D., A. Rushworth and G. Napier, "Spatio-temporal areal unit modelling in r with conditional autoregressive priors using the carbayesst package", Journal of Statistical Software (2017).

Legates, D. R. and C. J. Willmott, "Mean seasonal and spatial variability in gauge-corrected, global precipitation", International Journal of Climatology **10**, 2, 111–127 (1990).

Leroux, B. G., X. Lei and N. Breslow, "Estimation of disease rates in small areas: a new mixed model for spatial dependence", in "Statistical models in epidemiology, the environment, and clinical trials", pp. 179–191 (Springer, 2000).

Lettenmaier, D. P., "Detection of trends in water quality data from records with dependent observations", Water Resources Research **12**, 5, 1037–1046 (1976).

Levis, S., G. B. Bonan, E. Kluzek, P. E. Thornton, A. Jones, W. J. Sacks and C. J. Kucharik, "Interactive crop management in the community earth system model (cesm1): Seasonal influences on land–atmosphere fluxes", Journal of Climate **25**, 14, 4839–4859 (2012).

Li, G., N. Best, A. L. Hansell, I. Ahmed and S. Richardson, "Baystdetect: detecting unusual temporal patterns in small area data via bayesian model choice", Biostatistics **13**, 4, 695–710 (2012).

Liebmann, B., R. M. Dole, C. Jones, I. Bladé and D. Allured, "Influence of choice of time period on global surface temperature trend estimates", Bulletin of the American Meteorological Society **91**, 11, 1485–1492 (2010).

Lim, K.-S. S. and S.-Y. Hong, "Development of an effective double-moment cloud microphysics scheme with prognostic cloud condensation nuclei (ccn) for weather and climate models", Monthly weather review **138**, 5, 1587–1612 (2010).

Lloyd, C., *Spatial data analysis: an introduction for GIS users* (Oxford university press, 2010).

López-Bellido, L., J. Wery and R. J. López-Bellido, "Energy crops: prospects in the context of sustainable agriculture", European journal of agronomy **60**, 1–12 (2014).

López-Quilez, A. and F. Munoz, "Review of spatio-temporal models for disease mapping", Final Report for the EUROHEIS **2** (2009).

Mahmood, R., R. A. Pielke Sr, K. G. Hubbard, D. Niyogi, G. Bonan, P. Lawrence, R. McNider, C. McAlpine, A. Etter, S. Gameda *et al.*, "Impacts of land use/land cover change on climate and future research priorities", Bulletin of the American Meteorological Society **91**, 1, 37–46 (2010).

Mann, H. B., "Nonparametric tests against trend", Econometrica: Journal of the Econometric Society pp. 245–259 (1945).

McIsaac, G. F., M. B. David and C. A. Mitchell, "Miscanthus and switchgrass production in central illinois: impacts on hydrology and inorganic nitrogen leaching", Journal of environmental quality **39**, 5, 1790–1799 (2010).

McMillen, D. P., "Geographically weighted regression: the analysis of spatially varying relationships", (2004).

Meinshausen, N. and P. Bühlmann, "High-dimensional graphs and variable selection with the lasso", The annals of statistics pp. 1436–1462 (2006).

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, "Equation of state calculations by fast computing machines", The journal of chemical physics **21**, 6, 1087–1092 (1953).

Meyer, S., L. Held and M. Höhle, "Spatio-temporal analysis of epidemic phenomena using the r package surveillance", arXiv preprint arXiv:1411.0416 (2014).

Miguez, F. E., M. B. Villamil, S. P. Long and G. A. Bollero, "Meta-analysis of the effects of management factors on miscanthus× giganteus growth and biomass production", agricultural and forest meteorology **148**, 8-9, 1280–1292 (2008).

Miguez-Macho, G., G. L. Stenchikov and A. Robock, "Spectral nudging to eliminate the effects of domain position and geometry in regional climate model simulations", Journal of Geophysical Research: Atmospheres **109**, D13 (2004).

Miller, J. N., A. VanLoocke, N. Gomez-Casanovas and C. J. Bernacchi, "Candidate perennial bioenergy grasses have a higher albedo than annual row crops", Gcb Bioenergy **8**, 4, 818–825 (2016).

Moran, P. A., "The interpretation of statistical maps", Journal of the Royal Statistical Society. Series B (Methodological) **10**, 2, 243–251 (1948).

Mudelsee, M., *Climate time series analysis* (Springer, 2013).

Murakami, D. and D. A. Griffith, "Random effects specifications in eigenvector spatial filtering: a simulation study", Journal of Geographical Systems **17**, 4, 311–331 (2015).

Murakami, D. and D. A. Griffith, "Eigenvector spatial filtering for large data sets: fixed and random effects approaches", arXiv preprint arXiv:1702.06220 (2017).

Murakami, D., T. Yoshida, H. Seya, D. A. Griffith and Y. Yamagata, "A moran coefficient-based mixed effects approach to investigate spatially varying relationships", Spatial Statistics **19**, 68–89 (2017).

Murphy, L. N., W. J. Riley and W. D. Collins, "Local and remote climate impacts from expansion of woody biomass for bioenergy feedstock in the southeastern united states", Journal of Climate **25**, 21, 7643–7659 (2012).

NCEP, F., "Operational model global tropospheric analyses", Continuing From July (1999).

Noguchi, K., Y. R. Gel and C. R. Duguay, "Bootstrap-based tests for trends in hydrological time series, with application to ice phenology data", Journal of Hydrology **410**, 3-4, 150–161 (2011).

Oikawa, P. Y., G. D. Jenerette and D. A. Grantz, "Offsetting high water demands with high productivity: Sorghum as a biofuel crop in a high irradiance arid ecosystem", Gcb Bioenergy **7**, 5, 974–983 (2015).

Okalebo, J. A., R. J. Oglesby, S. Feng, K. Hubbard, A. Kilic, M. Hayes and C. Hays, "An evaluation of the community land model (version 3.5) and noah land surface models for temperature and precipitation over nebraska (central great plains): Implications for agriculture in simulations of future climate change and adaptation", in "Climate Change Adaptation, Resilience and Hazards", pp. 21–34 (Springer, 2016).

Pace, R. K., J. P. LeSage and S. Zhu, "Interpretation and computation of estimates from regression models using spatial filtering", Spatial Economic Analysis **8**, 3, 352–369 (2013).

Paciorek, C. J., "The importance of scale for spatial-confounding bias and precision of spatial regression estimators", Statistical science: a review journal of the Institute of Mathematical Statistics **25**, 1, 107 (2010).

Patuelli, R., D. A. Griffith, M. Tiefelsdorf and P. Nijkamp, "Spatial filtering and eigenvector stability: space-time models for german unemployment data", International Regional Science Review **34**, 2, 253–280 (2011).

Perlack, R. D., L. M. Eaton, A. F. Turhollow Jr, M. H. Langholtz, C. C. Brandt, M. E. Downing, R. L. Graham, L. L. Wright, J. M. Kavkewitz, A. M. Shamey *et al.*, "Us billion-ton update: biomass supply for a bioenergy and bioproducts industry", (2011).

Pielke, R. A., "Land use and climate change", Science **310**, 5754, 1625–1626 (2005).

Pinheiro, J., "nlme: Linear and nonlinear mixed effects models", http://cran. r-project. org/web/packages/nlme/index. html (2009).

Reich, B. J., J. S. Hodges and V. Zadnik, "Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models", Biometrics **62**, 4, 1197–1206 (2006).

Rue, H., S. Martino and N. Chopin, "Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations", Journal of the royal statistical society: Series b (statistical methodology) **71**, 2, 319–392 (2009).

Rushworth, A., D. Lee and R. Mitchell, "A spatio-temporal model for estimating the long-term effects of air pollution on respiratory hospital admissions in greater london", Spatial and spatio-temporal epidemiology **10**, 29–38 (2014).

Salamanca, F., M. Georgescu, A. Mahalov, M. Moustaoui and M. Wang, "Anthropogenic heating of the urban environment due to air conditioning", Journal of Geophysical Research: Atmospheres **119**, 10, 5949–5965 (2014).

Salamanca, F., M. Georgescu, A. Mahalov, M. Moustaoui, M. Wang and B. Svoma, "Assessing summertime urban air conditioning consumption in a semiarid environment", Environmental Research Letters **8**, 3, 034022 (2013).

Sansom, P. G., D. B. Stephenson, C. A. Ferro, G. Zappa and L. Shaffrey, "Simple uncertainty frameworks for selecting weighting schemes and interpreting multimodel ensemble climate change experiments", Journal of Climate **26**, 12, 4017–4037 (2013).

Sayemuzzaman, M. and M. K. Jha, "Seasonal and annual precipitation time series trend analysis in north carolina, united states", Atmospheric Research **137**, 183–194 (2014).

Schwarz, G. *et al.*, "Estimating the dimension of a model", The annals of statistics **6**, 2, 461–464 (1978).

Seya, H., D. Murakami, M. Tsutsumi and Y. Yamagata, "Application of lasso to the eigenvector selection problem in eigenvector-based spatial filtering", Geographical Analysis **47**, 3, 284–299 (2015).

Shen, X. and J. Ye, "Adaptive model selection", Journal of the American Statistical Association **97**, 457, 210–221 (2002).

Sigrist, F., H. R. Künsch, W. A. Stahel *et al.*, "spate: An r package for spatio-temporal modeling with a stochastic advection-diffusion process", Journal of Statistical Software **63**, 14, 1–23 (2015).

Skamarock, W. C., "A description of theadvanced research wrf version 3", Tech. Note pp. 1–96 (2008).

Smith, C. M., M. B. David, C. A. Mitchell, M. D. Masters, K. J. Anderson-Teixeira, C. J. Bernacchi and E. H. DeLucia, "Reduced nitrogen losses after conversion of row crop agriculture to perennial biofuel crops", Journal of environmental quality **42**, 1, 219–228 (2013).

Sonali, P. and D. N. Kumar, "Review of trend detection methods and their application to detect temperature changes in india", Journal of Hydrology **476**, 212–227 (2013).

Spiegelhalter, D. J., N. G. Best, B. P. Carlin and A. Van Der Linde, "Bayesian measures of model complexity and fit", Journal of the Royal Statistical Society: Series B (Statistical Methodology) **64**, 4, 583–639 (2002).

Taylor, K. E., "Summarizing multiple aspects of model performance in a single diagram", Journal of Geophysical Research: Atmospheres **106**, D7, 7183–7192 (2001).

Thayn, J. B. and J. M. Simanis, "Accounting for spatial autocorrelation in linear regression models using spatial filtering with eigenvectors", Annals of the Association of American Geographers **103**, 1, 47–66 (2013).

Tibshirani, R., "Regression shrinkage and selection via the lasso", Journal of the Royal Statistical Society. Series B (Methodological) pp. 267–288 (1996).

Tiefelsdorf, M. and D. A. Griffith, "Semiparametric filtering of spatial autocorrelation: the eigenvector approach", Environment and Planning A **39**, 5, 1193–1221 (2007).

Tobler, W. R., "A computer movie simulating urban growth in the detroit region", Economic geography **46**, sup1, 234–240 (1970).

Vanloocke, A., C. J. Bernacchi and T. E. Twine, "The impacts of miscanthus× giganteus production on the midwest us hydrologic cycle", Gcb Bioenergy **2**, 4, 180–191 (2010).

VanLoocke, A., T. E. Twine, C. J. Kucharik and C. J. Bernacchi, "Assessing the potential to decrease the gulf of mexico hypoxic zone with midwest us perennial cellulosic feedstock production", Gcb Bioenergy **9**, 5, 858–875 (2017).

Vogelsang, T. J., "Trend function hypothesis testing in the presence of serial correlation", Econometrica pp. 123–148 (1998).

von Storch, H., H. Langenberg and F. Feser, "A spectral nudging technique for dynamical downscaling purposes", Monthly weather review **128**, 10, 3664–3673 (2000).

Wagle, P. and V. G. Kakani, "Seasonal variability in net ecosystem carbon dioxide exchange over a young switchgrass stand", Gcb Bioenergy **6**, 4, 339–350 (2014).

Wagle, P., V. G. Kakani and R. L. Huhnke, "Evapotranspiration and ecosystem water use efficiency of switchgrass and high biomass sorghum", Agronomy Journal **108**, 3, 1007–1019 (2016).

Wagner, M., M. Wang, G. Miguez-Macho, J. Miller, A. VanLoocke, J. E. Bagley, C. J. Bernacchi and M. Georgescu, "A realistic meteorological assessment of perennial biofuel crop deployment: a southern great plains perspective", GCB Bioenergy **9**, 6, 1024–1041 (2017).

Wakefield, J., "Sensitivity analyses for ecological regression", Biometrics **59**, 1, 9–17 (2003).

Wang, H. and C. Leng, "Unified lasso estimation by least squares approximation", Journal of the American Statistical Association **102**, 479, 1039–1048 (2007).

Wang, M., M. Wagner, G. Miguez-Macho, Y. Kamarianakis, A. Mahalov, M. Moustaoui, J. Miller, A. VanLoocke, J. Bagley, C. Bernacchi *et al.*, "On the long-term hydroclimatic sustainability of perennial bioenergy crop expansion over the united states", Journal of Climate **30**, 7, 2535–2557 (2017).

Weaver, C. P. and R. Avissar, "Atmospheric disturbances caused by human modification of the landscape", Bulletin of the American Meteorological Society **82**, 2, 269–281 (2001).

Whittingham, M. J., P. A. Stephens, R. B. Bradbury and R. P. Freckleton, "Why do we still use stepwise modelling in ecology and behaviour?", Journal of animal ecology **75**, 5, 1182–1189 (2006).

Willmott, C. J. and K. Matsuura, "Smart interpolation of annually averaged air temperature in the united states", Journal of Applied Meteorology **34**, 12, 2577–2586 (1995).

Wilson, H., R. Cruse and C. Burras, "Perennial grass management impacts on runoff and sediment export from vegetated channels in pulse flow runoff events", biomass and bioenergy **35**, 1, 429–436 (2011).

Yue, S. and C. Wang, "The mann-kendall test modified by effective sample size to detect trend in serially correlated hydrological series", Water resources management **18**, 3, 201–218 (2004).

Zhu, P., Q. Zhuang, J. Eva and C. Bernacchi, "Importance of biophysical effects on climate warming mitigation potential of biofuel crops over the conterminous united states", Gcb Bioenergy **9**, 3, 577–590 (2017).

Zhuang, Q., Z. Qin and M. Chen, "Biofuel, land and water: maize, switchgrass or miscanthus?", Environmental Research Letters **8**, 1, 015020 (2013).

Zou, H., "The adaptive lasso and its oracle properties", Journal of the American statistical association **101**, 476, 1418–1429 (2006).

APPENDIX A

SUPPLEMENTAL MATERIAL FOR CHAPTER 2

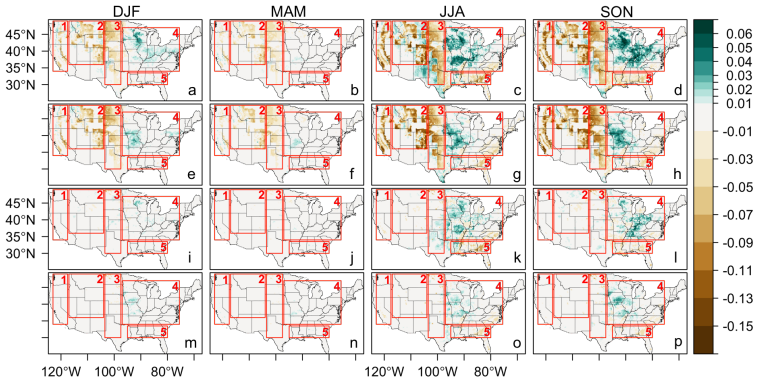

Figure A.1: Seasonally averaged soil moisture difference (m3 m-3) at 10-40 cm soil depth (Perennial100_E1-Control_E1) over one decade (2000-2009) for (a) DJF, (b) MAM, (c) JJA, and (d) SON. (e)-(h) Same as (a)-(d) but for difference of Perennial100_E8 minus Control_E8. (i)-(l) Same as (a)-(d) but for difference of Perennial25_E1 minus Control_E1. (m)-(p) Same as (a)-(d) but for difference of Perennial25_E8 minus Control_E8. Red rectangles outline five sub-regions for time series calculations.

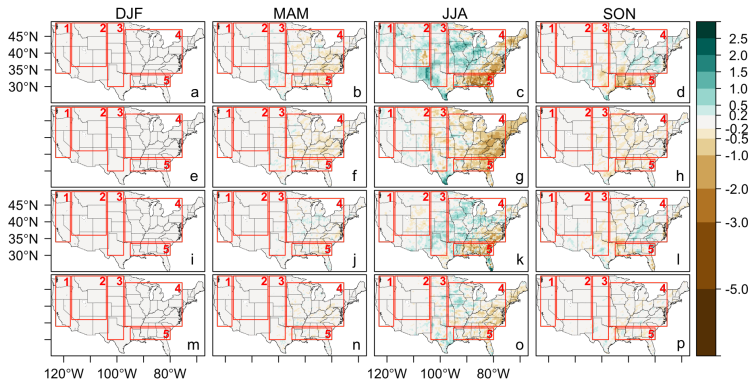

Figure A.2: Seasonally averaged precipitation difference (mm d-1) (Perennial100_E1-Control_E1) over one decade (2000-2009) for (a) DJF, (b) MAM, (c) JJA, and (d) SON. (e)-(h) Same as (a)-(d) but for difference of Perennial100_E8 minus Control_E8. (i)-(l) Same as (a)-(d) but for difference of Perennial25_E1 minus Control_E1. (m)-(p) Same as (a)-(d) but for difference of Perennial25_E8 minus Control_E8. Red rectangles outline five sub-regions for time series calculations.

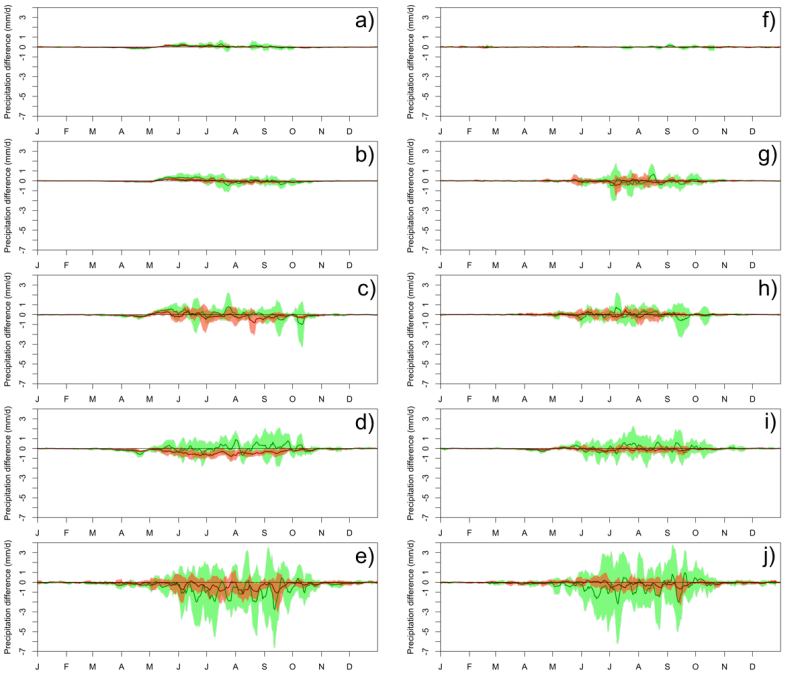

Figure A.3: Annual cycle of precipitation difference (mm d-1) averaged only over grid cells undergoing land surface modification under Perennial100 scenario (a) region 1, (b) region 2, (c) region 3, (d) region 4 and (e) region 5. (f)-(j) Same as (a)-(e) but under Perennial25 scenario. Green and red lines indicate averaged annual cycle of simulated impact over decadal period using ensemble member E1 and E8, respectively. Bands of one standard deviation above and below the mean annual cycle are shaded with the corresponding color.

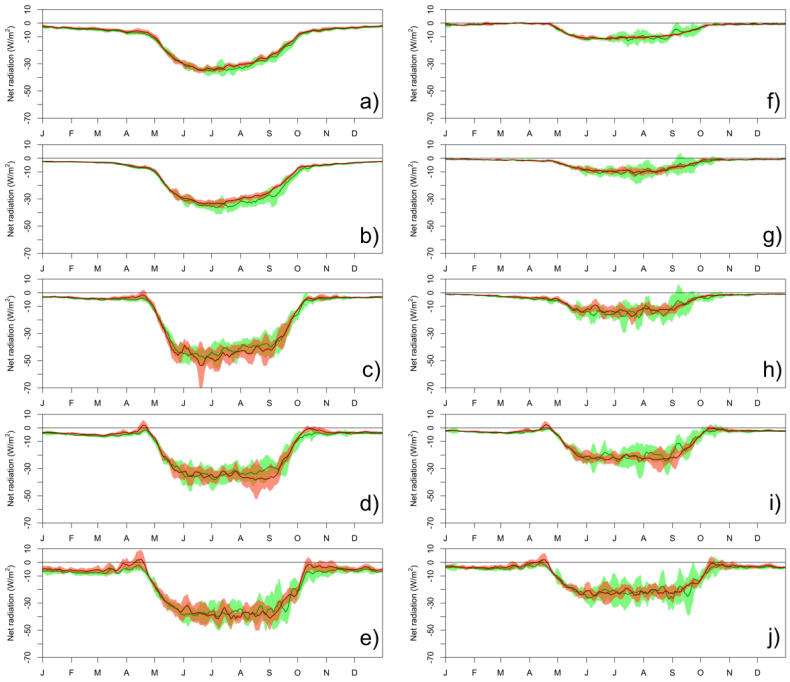

Figure A.4: Annual cycle of net radiation difference (W m-2) averaged only over grid cells undergoing land surface modification under Perennial100 scenario (a) region 1, (b) region 2, (c) region 3, (d) region 4, and (e) region 5. (f)-(j) Same as (a)-(e) but under Perennial25 scenario. Green and red lines indicate averaged annual cycle of simulated impact over decadal period using ensemble member E1 and E8, respectively. Bands of one standard deviation above and below the mean annual cycle are shaded with the corresponding color.

APPENDIX B

R CODE FOR CHAPTER 4

B.1 Data generating process for Scenario EFM1 and EFMX

```
rm(list=ls())
source("~/STEF_function.R")


library(raster)
library(ngspatial)
library(MASS)
library(monomvn)
library(spdep)
library(orcutt)
library(CARBayesST)
library(glmnet)
library(parcor)
library(lsa)
library(genlasso)
library(Matrix)
library(demogR)
library(horseshoe)
library(CompRandFld)
library(spdep)

#############################################################################
t=20 #20 time points
nlon=10 #number of lat points
nlat=10 #number of lon points
ntotal=nlon*nlat #number of total pixels
ntotal_st = ntotal*t #number of total pixels*time
rep=50  #number of replicates of spatial-temporal datasets
sd_val = c(0.001,1,10)
```

```
#############################################################################
#create spatial-temporal pixels and get coordinates
#############################################################################
#create spatial-temporal brick
r <- brick(nrow=nlat, ncol=nlon,nl=t,xmn=0.05, xmx=1.05,ymn=0.05,ymx=1.05)
#get coordinates
x = coordinates(r)
# big X matrix in space and time
X_st_big = replicate(t,x) #lon,lat
X_st_big_vec = apply(X_st_big, 2, c) #big X matrix

#############################################################################
#create spatial contiguity matrix based on pixel
#############################################################################
#raster to polygon
r_poly=rasterToPolygons(r[[1]], fun=NULL, n=4, na.rm=TRUE, digits=12,
    dissolve=FALSE)
plot(r_poly)
n1 =poly2nb(r_poly, queen=TRUE) #Construct neighbours list from polygon list
A <- nb2mat(n1,style="B") #Spatial weights matrices for neighbours lists

#############################################################################
#temporal neighboring matrix
#############################################################################
A_t=odiag(rep(1,t-1),-1)+odiag(rep(1,t-1), 1)
A_t_raster=raster(A_t)
extent(A_t_raster) = c(0.5,20.5,0.5,20.5)
A_t_polygon=rasterToPolygons(A_t_raster)

#############################################################################
#space-time adjacency matrix(contemporaneous)
#############################################################################
A_contemp = kronecker(diag(1,t),A)+kronecker(A_t,diag(1,ntotal))

#############################################################################
#Moran operator for creating y
#############################################################################
```

```
#For Scenario EFMX:
P_st = X_st_big_vec %*% solve(t(X_st_big_vec) %*% X_st_big_vec) %*%
    t(X_st_big_vec)
#For Scenario EFM1:
#P_st = rep(1,ntotal_st)%*%t(rep(1,ntotal_st))/ntotal_st

P_st_orthogonal = diag(ntotal_st)-P_st
Moran_op_contemp = P_st_orthogonal %*% A_contemp %*% P_st_orthogonal
eigen_Moran = eigen(Moran_op_contemp)

#check positive eigenvalues
is.zero = function(x, tol = .Machine$double.eps^0.5)
{abs(x) < tol}
eigenvalues=eigen_Moran$values
maxatt = match(TRUE, sapply(eigenvalues, is.zero)) - 1
print(maxatt)

#eigenvectors of Moran operator
M_contemp_positive = eigen(Moran_op_contemp)$vectors[,1:maxatt]
# take eigenvectors w.r.t all positive eigenvalues for estimating

#only take the first 200 eigenvectors of Moran operator for generating data
n_M_vector = 200
M_contemp = eigen_Moran$vectors[,1:n_M_vector]

##############################################################################
#generating values for y
##############################################################################
#coefficients for lat and lon
beta=c(1,1)
# make values 1 to 10, with each rep 5 times
delta_s_values = rep(seq(1,10),each=5)
Y_EF_all=list(list(),list(),list()) #store all spatial-temporal dependent
    datasets
delta_s_all=list(list(),list(),list()) #store delta

for (s in 1:3){ #loop for sigma
  cat("s = ", s, "\n")
```

```
  for(r in 1:rep){ #loop for replicates 50 times
    cat("rep = ", r, "\n")

    delta_s_all[[s]][[r]] = rep(0,n_M_vector)
    location = sample(1:200,length(delta_s_values))
    delta_s_all[[s]][[r]][location] = delta_s_values

    # spatio-temporal random component
    W = M_contemp%*%delta_s_all[[s]][[r]]

    # simulate obs
    Y_mean=X_st_big_vec%*%beta + W
    Y_EF_all[[s]][[r]] = mvrnorm(1,mu=Y_mean,Sigma=diag(ntotal_st)*sd_val[s])
  }
}
```

B.2 Data generating process for Scenario AR

```
###############################################################################
t=20 #20 time points
nlon=10 #number of lat points
nlat=10 #number of lon points
ntotal=nlon*nlat #number of total pixels
ntotal_st = ntotal*t #number of total pixels*time
rep=50  #number of replicates of spatial-temporal datasets
sd_val = c(0.001,1,10)

###############################################################################
#create spatial-temporal pixels and get coordinates
###############################################################################
#create spatial-temporal brick
r <- brick(nrow=nlat, ncol=nlon,nl=t,xmn=0.05, xmx=1.05,ymn=0.05,ymx=1.05)
#get coordinates
x = coordinates(r)
# big X matrix in space and time
X_st_big = replicate(t,x) #lon,lat
X_st_big_vec = apply(X_st_big, 2, c) #big X matrix rbind
```

```
#############################################################################
#create spatial contiguity matrix based on pixel
#############################################################################
#raster to polygon
r_poly=rasterToPolygons(r[[1]], fun=NULL, n=4, na.rm=TRUE, digits=12,
    dissolve=FALSE)
plot(r_poly)
#polygon to nb
n1 =poly2nb(r_poly, queen=TRUE) #Construct neighbours list from polygon list
A <- nb2mat(n1,style="B") #Spatial weights matrices for neighbours lists

#############################################################################
#temporal neighboring matrix
#############################################################################
A_t=odiag(rep(1,t-1),-1)+odiag(rep(1,t-1), 1)
A_t_raster=raster(A_t)
extent(A_t_raster) = c(0.5,20.5,0.5,20.5)
A_t_polygon=rasterToPolygons(A_t_raster)

#############################################################################
# generate rho values of AR in space
#############################################################################
rho = apply(x, 1, max)/1.2
# create a raster for rho, only for plot
rho_raster <- r[[1]]
rho_raster[] <- rho   #note the different pixel index

#############################################################################
# generate rho values of AR in space
#############################################################################
beta=c(1,1)  #define coefficients for x (lat&lon)

#############################################################################
# generate values for y
#############################################################################
  Y_AR_all=list(list(),list(),list())
  for (s in 1:length(sd_val)){ # loop for standard deviation values
```

```
    cat("s = ", s, "\n")
    for (r in 1:rep){ #loop for replicates 50 times
      cat("rep = ", r, "\n")
      Y_AR_mtx=matrix(NA,nrow=ntotal,ncol=t)
      for (i in 1:ntotal){ #loop for pixel
        cat("pixel = ", i, "\n")
        Y_AR_mtx[i,]=arima.sim(model=list(ar=rho[i]),sd=sd_val[s],n=t) #random
            component
      }
      Y_AR_all[[s]][[r]]=X_st_big_vec%*%beta+as.vector(Y_AR_mtx)
    }
  }
```

B.3 Data generating process for Scenario RF

```
##############################################################################
t=20 #20 time points
nlon=10 #number of lat points
nlat=10 #number of lon points
ntotal=nlon*nlat #number of total pixels
ntotal_st = ntotal*t #number of total pixels*time
rep=50  #number of replicates of spatial-temporal datasets

#create spatial-temporal brick
r <- brick(nrow=nlat, ncol=nlon,nl=t,xmn=0.05, xmx=1.05,ymn=0.05,ymx=1.05)
#raster to polygon
r_poly=rasterToPolygons(r[[1]], fun=NULL, n=4, na.rm=TRUE, digits=12,
    dissolve=FALSE)
plot(r_poly)
##############################################################################
# create spatio-temporal gaussian random field
##############################################################################
# with separable spatio-temporal process, covariance exp_exp
# Define the spatial-coordinates of the points:
x <- seq(0.1, 1, 0.1)
y <- seq(0.1, 1, 0.1)
# Define the temporal-coordinates:
times <- seq(1, t, 1) #20 years
```

```
#############################################################################
# create spatio-temporal gaussian random field
#with nonseparable spatio-temporal process, covariance gneiting
#############################################################################
  #check parameter of correlation matrix:
  CorrelationParam("gneiting")  #"power_s" "power_t" "scale_s" "scale_t" "sep"
  #define values of parameters

  data_gneiting=list() #store all spatial-temporal dependent datasets
  for(r in 1:rep){ #loop for replicates 10 times
    cat("rep = ", r, "\n")
    data_gneiting[[r]] <- RFsim(x, y, times, corrmodel="gneiting",grid=TRUE,
                                param=list(mean = 0, nugget = 0, sill = 1,
                                    power_s = 1,
                                        power_t = 1.544, scale_s = 0.00134,
                                            scale_t = 0.901, sep = 1))$data
  }

#############################################################################
#create spatial-temporal pixels and get coordinates
#############################################################################
#create spatial-temporal brick
r <- brick(nrow=nlat, ncol=nlon,nl=t,xmn=0.05, xmx=1.05,ymn=0.05,ymx=1.05)
#get coordinates
x = coordinates(r)
# big X matrix in space and time
X_st_big = replicate(t,x) #lon,lat
X_st_big_vec = apply(X_st_big, 2, c) #big X matrix rbind

#############################################################################
#create spatial contiguity matrix based on pixel
#############################################################################
#raster to polygon
r_poly=rasterToPolygons(r[[1]], fun=NULL, n=4, na.rm=TRUE, digits=12,
    dissolve=FALSE)
plot(r_poly)
#polygon to nb
```

```
#library(spdep)
n1 =poly2nb(r_poly, queen=TRUE) #Construct neighbours list from polygon list
A <- nb2mat(n1,style="B") #Spatial weights matrices for neighbours lists

##############################################################################
#temporal neighboring matrix
##############################################################################
A_t=odiag(rep(1,t-1),-1)+odiag(rep(1,t-1), 1)
plot(raster(A_t)) #check neighboring matrix

##############################################################################
# generate values for y
##############################################################################
#coefficients for lat and lon
beta=c(1,1)  #define coefficients for x (lat&lon)

  Y_RF=list() #store all spatial-temporal dependent datasets
  for(r in 1:rep){ #loop for replicates
    cat("rep = ", r, "\n")
    Y_RF[[r]]=X_st_big_vec%*%beta + as.vector(data_gneiting[[r]])
  }
```

B.4 STEF algorithms

```
##############################################################################
#space-time adjacency matrix(contemporaneous)
A_contemp = kronecker(diag(1,t),A)+kronecker(A_t,diag(1,ntotal))


##############################################################################
#For method NCF
##############################################################################
#Moran operator
##############################################################################
P_st = X_st_big_vec %*% solve(t(X_st_big_vec) %*% X_st_big_vec) %*%
    t(X_st_big_vec)

P_st = rep(1,ntotal_st)%*%t(rep(1,ntotal_st))/ntotal_st
```

```
P_st_orthogonal = diag(ntotal_st)-P_st
Moran_op_contemp = P_st_orthogonal %*% A_contemp %*% P_st_orthogonal
eigen_Moran = eigen(Moran_op_contemp)

#check positive eigenvalues
is.zero = function(x, tol = .Machine$double.eps^0.5)
{abs(x) < tol}
eigenvalues=eigen_Moran$values
maxatt = match(TRUE, sapply(eigenvalues, is.zero)) - 1
print(maxatt)

#eigenvectors of Moran operator
M_contemp_positive = eigen_Moran$vectors[,1:maxatt]
# take eigenvectors w.r.t all positive eigenvalues for estimating

# design matrix of lat,lon,eigenvectors
X_design_contemp = cbind(X_st_big_vec,M_contemp_positive)



##############################################################################################
#For method CF
##############################################################################################
#Moran operator
##############################################################################################
P_st = rep(1,ntotal_st)%*%t(rep(1,ntotal_st))/ntotal_st
P_st_orthogonal = diag(ntotal_st)-P_st
Moran_op_contemp = P_st_orthogonal %*% A_contemp %*% P_st_orthogonal
eigen_Moran = eigen(Moran_op_contemp)

#check positive eigenvalues
is.zero = function(x, tol = .Machine$double.eps^0.5)
{abs(x) < tol}
eigenvalues=eigen_Moran$values
maxatt = match(TRUE, sapply(eigenvalues, is.zero)) - 1
print(maxatt)

#eigenvectors of Moran operator
```

```
M_contemp_positive = eigen_Moran$vectors[,1:maxatt]
# take eigenvectors w.r.t all positive eigenvalues for estimating

# design matrix of lat,lon,eigenvectors
X_design_contemp = cbind(X_st_big_vec,M_contemp_positive)




##############################################################################################
#For method ICF
##############################################################################################
#Moran operator
##############################################################################################
P_st = rep(1,ntotal_st)%*%t(rep(1,ntotal_st))/ntotal_st
P_st_orthogonal = diag(ntotal_st)-P_st
Moran_op_contemp = P_st_orthogonal %*% A_contemp %*% P_st_orthogonal
eigen_Moran = eigen(Moran_op_contemp)

##############################################################################################
#take all eigenvectors
M_contemp = eigen_Moran$vectors
# design matrix of lat,lon,eigenvectors
X_design_contemp = cbind(X_st_big_vec,M_contemp)

##############################################################################################
# apply VIF on explanatory variables
##############################################################################################
X_st_big_vec_Xvif = vif_func(X_st_big_vec,thresh=10,trace=T)

##############################################################################################
# apply VIF on eigenvectors
##############################################################################################
M_contemp_VIF = STEF_vif_func(in_frame = M_contemp, X = X_st_big_vec_Xvif)

##############################################################################################
#apply SIS
##############################################################################################
```

```
proc_time_SIS = list(list(),list(),list())
M_contemp_VIF_SIS = list(list(),list(),list())
M_contemp_VIF_SIS_eff = list(list(),list(),list())
d = ceiling(ntotal_st/log(ntotal_st))
d_eff = list(list(),list(),list())

for (s in 1:length(sd_val)){ # loop for standard deviation values
  cat("s = ", s, "\n")

  for(r in 1:rep){ #loop for replicates
    cat("rep = ", r, "\n")

    cor_result =
        cbind(1:dim(M_contemp_VIF[[1]])[2],cor(M_contemp_VIF[[1]],Y_EF_all[[s]][[r]]))
    cor_result_order = cor_result[order(cor_result[,2],decreasing=TRUE),]

    M_contemp_VIF_SIS[[s]][[r]] = M_contemp_VIF[[1]][,cor_result_order[1:d]]

    #get effective sample size
    res = lm(Y_EF_all[[s]][[r]]~X_st_big_vec_Xvif-1)$residuals   # get residuals
    reg_vif = STEF_VIF(lm(res~M_contemp_VIF_SIS[[s]][[r]]-1))
    ntotal_st_eff = 1/reg_vif * ntotal_st

    #get number of selected eigenvectors by SIS
    d_eff[[s]][[r]] = ceiling(ntotal_st_eff/log(ntotal_st_eff))

    M_contemp_VIF_SIS_eff[[s]][[r]] =
        M_contemp_VIF[[1]][,cor_result_order[1:d_eff[[s]][[r]],1]]
  }
}
```

B.5 STEF functions

```
source("LSA.R")

##############################################################################
#calculate VIF
##############################################################################
```

```
STEF_VIF = function (X)
{
  # X is a lm object
  1/(1 - summary(X)$r.squared)
}

#############################################################################
# VIF filtering for explanatory variables
#it applied a stepwise procedure to remove predictors
#############################################################################
vif_func<-function(in_frame,thresh=10,trace=T){

  in_frame = X_st_big_vec

  require(fmsb)

  if(class(in_frame) != "data.frame") in_frame<-data.frame(in_frame)

  #get initial vif value for all comparisons of variables
  vif_init<-NULL
  for(val in names(in_frame)){
    form_in<-formula(paste(val," ~ ."))
    vif_init<-rbind(vif_init,c(val,VIF(lm(form_in,data=in_frame))))
  }
  vif_max<-max(as.numeric(vif_init[,2]))

  if(vif_max < thresh){
    if(trace==T){ #print output of each iteration
      prmatrix(vif_init,collab=c("var","vif"),rowlab=rep("",nrow(vif_init)),quote=F)
      cat("\n")
      cat(paste("All variables have VIF < ", thresh,", max VIF
          ",round(vif_max,2), sep=""),"\n\n")
    }
    return(as.matrix(in_frame))
  }

  else{
```

```
    in_dat<-in_frame

    #backwards selection of explanatory variables, stops when all VIF values
        are below "thresh"
    while(vif_max >= thresh){

      vif_vals<-NULL

      for(val in names(in_dat)){
        form_in<-formula(paste(val," ~ ."))
        vif_add<-VIF(lm(form_in,data=in_dat))
        vif_vals<-rbind(vif_vals,c(val,vif_add))
      }
      max_row<-which(vif_vals[,2] == max(as.numeric(vif_vals[,2])))[1]

      vif_max<-as.numeric(vif_vals[max_row,2])

      if(vif_max<thresh) break

      if(trace==T){ #print output of each iteration
        # this is an older version of print.matrix
        prmatrix(vif_vals,collab=c("var","vif"),rowlab=rep("",nrow(vif_vals)),quote=F)
        cat("\n")
        cat("removed: ",vif_vals[max_row,1],vif_max,"\n\n")
        flush.console()
      }

      in_dat<-in_dat[,!names(in_dat) %in% vif_vals[max_row,1]]
    }

    return(as.matrix(in_dat))
  }
}
```

```
##############################################################################
```

```
# VIF filtering for filtering eigenvectors
#############################################################################
STEF_vif_func<-function(in_frame,X,thresh=10,trace=T){
  #in_frame is matrix of eigenvectors; X is explanatory variables
  #get initial vif value for all comparisons of variables
  vif_init<-NULL
  for(val in 1:dim(in_frame)[2]){
    testdata = cbind(in_frame[,val],X)
    vif_init<-rbind(vif_init,c(val,STEF_VIF(lm(testdata[,1]~testdata[,-1]))))
  }

  vif_init = vif_init[vif_init[,2]!=Inf,]

  large_row<-which(vif_init[,2] >= thresh)
  vif_max<-max(as.numeric(vif_init[,2]))

  if(vif_max < thresh){
    if(trace==T){ #print output of each iteration
      prmatrix(vif_init,collab=c("var","vif"),rowlab=rep("",nrow(vif_init)),quote=F)
      cat("\n")
      cat(paste("All variables have VIF < ", thresh,", max VIF
          ",round(vif_max,2), sep=""),"\n\n")
    }
    in_dat = in_frame
  }

  else{
    in_dat<-in_dat[,-large_row]
  }
  return(list(in_dat,large_row,vif_init))
}



#############################################################################
#method 1
#############################################################################
#function
```

```
STEF_method1 = function(Y,X,M_contemp_positive){

  #direct LSA,regress y on x & all eigenvectors
  X_design_contemp = cbind(X,M_contemp_positive)
  lm_out <- lm(Y ~ X_design_contemp-1)
  lsa_out <- lsa(lm_out)
  lsa_coef_aic<-as.numeric(lsa_out$beta.aic)
  lsa_coef_bic<-as.numeric(lsa_out$beta.bic)

  #aic
  #get selected eigenvectors
  nonzero_aic = as.matrix(M_contemp_positive)[,(lsa_coef_aic[-c(1,2)]!=0)]
  #design matrix including x and selected eigenvectors
  X_nonzero_aic = cbind(X,nonzero_aic)
  #regress y on x & selected eigenvectors
  lm_out_final_aic <- lm(Y ~ X_nonzero_aic -1)

  #bic
  #get selected eigenvectors
  nonzero_bic = as.matrix(M_contemp_positive)[,(lsa_coef_bic[-c(1,2)]!=0)]
  #design matrix including x and selected eigenvectors
  X_nonzero_bic = cbind(X,nonzero_bic)
  #regress y on x & selected eigenvectors
  lm_out_final_bic <- lm(Y ~ X_nonzero_bic -1)


  return_list = list(lm_out,lsa_out,lsa_coef_aic,lsa_coef_bic,
                     nonzero_aic,X_nonzero_aic,lm_out_final_aic,
                     nonzero_bic,X_nonzero_bic,lm_out_final_bic,
                     proc_time_STEF)
  names(return_list) = c("lm_out","lsa_out","lsa_coef_aic","lsa_coef_bic",
                         "nonzero_aic","X_nonzero_aic","lm_out_final_aic",
                         "nonzero_bic","X_nonzero_bic","lm_out_final_bic")
  return(return_list)
}
```

```
#############################################################################
#method 2
#############################################################################
#function
STEF_method2 = function(Y,X,M_contemp_positive){

  #stage 1: conventional regression of y with latlon, get residuals
  lm_residual=lm(Y~X-1)$residuals
  # lsa on residuals
  lm_residual_out <- lm(lm_residual ~ M_contemp_positive -1)
  lsa_out <- lsa(lm_residual_out)

  #select eigenvectors
  nonzero_aic<-M_contemp_positive[,(lsa_out$beta.aic !=0)]
  nonzero_bic<-M_contemp_positive[,(lsa_out$beta.bic !=0)]

  # stage 2:re-estimate coef of x
  X_design_final_aic = cbind(X,nonzero_aic)
  X_design_final_bic = cbind(X,nonzero_bic)
  lm_out_final_aic=lm(Y~X_design_final_aic -1)
  lm_out_final_bic=lm(Y~X_design_final_bic -1)

  return_list = list(lm_residual,lm_residual_out,lsa_out,
                     nonzero_aic,nonzero_bic,
                     X_design_final_aic,X_design_final_bic,
                     lm_out_final_aic,lm_out_final_bic,
                     proc_time_STEF)
  names(return_list) = c("lm_residual","lm_residual_out","lsa_out",
                         "nonzero_aic","nonzero_bic",
                         "X_design_final_aic","X_design_final_bic",
                         "lm_out_final_aic","lm_out_final_bic")
  return(return_list)
}
```

B.6 STEF functions

```
source("LSA.R")
```

```
#############################################################################
#calculate VIF
#############################################################################
STEF_VIF = function (X)
{
  # X is a lm object
  1/(1 - summary(X)$r.squared)
}

#############################################################################
# VIF filtering for explanatory variables
#it applied a stepwise procedure to remove predictors
#############################################################################
vif_func<-function(in_frame,thresh=10,trace=T){

  in_frame = X_st_big_vec

  require(fmsb)

  if(class(in_frame) != "data.frame") in_frame<-data.frame(in_frame)

  #get initial vif value for all comparisons of variables
  vif_init<-NULL
  for(val in names(in_frame)){
    form_in<-formula(paste(val," ~ ."))
    vif_init<-rbind(vif_init,c(val,VIF(lm(form_in,data=in_frame))))
  }
  vif_max<-max(as.numeric(vif_init[,2]))

  if(vif_max < thresh){
    if(trace==T){ #print output of each iteration
      prmatrix(vif_init,collab=c("var","vif"),rowlab=rep("",nrow(vif_init)),quote=F)
      cat("\n")
      cat(paste("All variables have VIF < ", thresh,", max VIF
          ",round(vif_max,2), sep=""),"\n\n")
    }
    return(as.matrix(in_frame))
  }
```

```
  else{

    in_dat<-in_frame

    #backwards selection of explanatory variables, stops when all VIF values
        are below "thresh"
    while(vif_max >= thresh){

      vif_vals<-NULL

      for(val in names(in_dat)){
        form_in<-formula(paste(val," ~ ."))
        vif_add<-VIF(lm(form_in,data=in_dat))
        vif_vals<-rbind(vif_vals,c(val,vif_add))
      }
      max_row<-which(vif_vals[,2] == max(as.numeric(vif_vals[,2])))[1]

      vif_max<-as.numeric(vif_vals[max_row,2])

      if(vif_max<thresh) break

      if(trace==T){ #print output of each iteration
        # this is an older version of print.matrix
        prmatrix(vif_vals,collab=c("var","vif"),rowlab=rep("",nrow(vif_vals)),quote=F)
        cat("\n")
        cat("removed: ",vif_vals[max_row,1],vif_max,"\n\n")
        flush.console()
      }

      in_dat<-in_dat[,!names(in_dat) %in% vif_vals[max_row,1]]
    }

    return(as.matrix(in_dat))
  }
}
```

```
#############################################################################
# VIF filtering for filtering eigenvectors
#############################################################################
STEF_vif_func<-function(in_frame,X,thresh=10,trace=T){
  #in_frame is matrix of eigenvectors; X is explanatory variables
  #get initial vif value for all comparisons of variables
  vif_init<-NULL
  for(val in 1:dim(in_frame)[2]){
    testdata = cbind(in_frame[,val],X)
    vif_init<-rbind(vif_init,c(val,STEF_VIF(lm(testdata[,1]~testdata[,-1]))))
  }

  vif_init = vif_init[vif_init[,2]!=Inf,]

  large_row<-which(vif_init[,2] >= thresh)
  vif_max<-max(as.numeric(vif_init[,2]))

  if(vif_max < thresh){
    if(trace==T){ #print output of each iteration
      prmatrix(vif_init,collab=c("var","vif"),rowlab=rep("",nrow(vif_init)),quote=F)
      cat("\n")
      cat(paste("All variables have VIF < ", thresh,", max VIF
          ",round(vif_max,2), sep=""),"\n\n")
    }
    in_dat = in_frame
  }

  else{
    in_dat<-in_dat[,-large_row]
  }
  return(list(in_dat,large_row,vif_init))
}



#############################################################################
```

```
#method 1
#############################################################################
#function

STEF_method1 = function(Y,X,M_contemp_positive){

  #direct LSA,regress y on x & all eigenvectors
  X_design_contemp = cbind(X,M_contemp_positive)
  lm_out <- lm(Y ~ X_design_contemp-1)
  lsa_out <- lsa(lm_out)
  lsa_coef_aic<-as.numeric(lsa_out$beta.aic)
  lsa_coef_bic<-as.numeric(lsa_out$beta.bic)

  #aic
  #get selected eigenvectors
  nonzero_aic = as.matrix(M_contemp_positive)[,(lsa_coef_aic[-c(1,2)]!=0)]
  #design matrix including x and selected eigenvectors
  X_nonzero_aic = cbind(X,nonzero_aic)
  #regress y on x & selected eigenvectors
  lm_out_final_aic <- lm(Y ~ X_nonzero_aic -1)

  #bic
  #get selected eigenvectors
  nonzero_bic = as.matrix(M_contemp_positive)[,(lsa_coef_bic[-c(1,2)]!=0)]
  #design matrix including x and selected eigenvectors
  X_nonzero_bic = cbind(X,nonzero_bic)
  #regress y on x & selected eigenvectors
  lm_out_final_bic <- lm(Y ~ X_nonzero_bic -1)


  return_list = list(lm_out,lsa_out,lsa_coef_aic,lsa_coef_bic,
                     nonzero_aic,X_nonzero_aic,lm_out_final_aic,
                     nonzero_bic,X_nonzero_bic,lm_out_final_bic,
                     proc_time_STEF)
  names(return_list) = c("lm_out","lsa_out","lsa_coef_aic","lsa_coef_bic",
                         "nonzero_aic","X_nonzero_aic","lm_out_final_aic",
                         "nonzero_bic","X_nonzero_bic","lm_out_final_bic")
  return(return_list)
```

```
}


#############################################################################
#method 2
#############################################################################
#function
STEF_method2 = function(Y,X,M_contemp_positive){

  #stage 1: conventional regression of y with latlon, get residuals
  lm_residual=lm(Y~X-1)$residuals
  # lsa on residuals
  lm_residual_out <- lm(lm_residual ~ M_contemp_positive -1)
  lsa_out <- lsa(lm_residual_out)

  #select eigenvectors
  nonzero_aic<-M_contemp_positive[,(lsa_out$beta.aic !=0)]
  nonzero_bic<-M_contemp_positive[,(lsa_out$beta.bic !=0)]

  # stage 2:re-estimate coef of x
  X_design_final_aic = cbind(X,nonzero_aic)
  X_design_final_bic = cbind(X,nonzero_bic)
  lm_out_final_aic=lm(Y~X_design_final_aic -1)
  lm_out_final_bic=lm(Y~X_design_final_bic -1)

  return_list = list(lm_residual,lm_residual_out,lsa_out,
                     nonzero_aic,nonzero_bic,
                     X_design_final_aic,X_design_final_bic,
                     lm_out_final_aic,lm_out_final_bic,
                     proc_time_STEF)
  names(return_list) = c("lm_residual","lm_residual_out","lsa_out",
                         "nonzero_aic","nonzero_bic",
                         "X_design_final_aic","X_design_final_bic",
                         "lm_out_final_aic","lm_out_final_bic")
  return(return_list)
}
```