CYBERGIS-ENABLED SPATIAL DECISION SUPPORT FOR SUPPLY CHAIN OPTIMIZATION WITH UNCERTAINTY QUANTIFICATION

BY

HAO HU

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Geography
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2018

Urbana, Illinois

Doctoral Committee:

Professor Shaowen Wang, Chair
Associate Professor Bo Li
Associate Professor Luis F. Rodríguez
Professor Mei-Po Kwan
Professor Yanfeng Ouyang

ABSTRACT

Spatial decision support systems have made extensive progress on taking advantage of geographic information science and systems (GIS) for the synthesis of geospatial data and analysis, domain-specific knowledge and models, and advanced computing technologies. However, a major challenge revolving around the synthesis remains to systematically quantify uncertainties of complex data, models, and computation. For example, the state of the art of supply chain optimization does not adequately address uncertainty in the context of spatial decision support. This challenge is caused in part by the computational intensity of uncertainty quantification and propagation through optimization models.

This research aims to establish a novel cyberGIS framework for resolving the computational intensity to incorporate uncertainty quantification into spatial decision support. Specifically, the cyberGIS framework seamlessly integrates uncertainty quantification and supply chain optimization modeling into a CyberGIS Gateway application that represents a cutting-edge online cyberGIS environment for users to perform interactive spatial decision-making enabled by advanced cyberinfrastructure. Furthermore, an innovative method combining Bayesian hierarchical modeling with stochastic programming is proposed to explicitly account for spatiotemporal uncertainties in supply chain optimization. The cyberGIS framework and related method are evaluated based on a case study of the biomass-to-bioenergy supply chain optimization at the county level in the United States to resolve the synthesis challenge in multiple spatial decision support scenarios.

Keywords: spatial decision support, cyberGIS, uncertainty and sensitivity analysis, supply chain optimization, spatiotemporal data analysis

To my family with love

ACKNOWLEDGEMENT

The work leading toward the completion of my dissertation has been a journey unlike any other in my lifetime. This journey would not have been possible without the support and help from family, friends, and advisors.

I would first like to thank my advisor, Dr. Shaowen Wang. He has contributed so much to my growth throughout the years, not only on the improvement of research skills such as enabling me to think critically and teaching me how to be a researcher, but also on the transferable skills such as communication, collaboration, and project management skills. His passion for research inspired me and his dedication to scholarship will continue to be a model for me to follow into the future.

I gratefully acknowledge my Ph.D. committee members, Dr. Luis Rodríguez, Dr. Mei-Po Kwan, Dr. Bo Li, and Dr. Yanfeng Ouyang for serving on my doctoral committee and providing valuable feedback on my research. I am fortunate to have had the opportunity to work with them to gain multi-disciplinary knowledge. This dissertation could not have been possible without their support and guidance.

I owe gratitude to many people at the University of Illinois at Urbana-Champaign (UIUC) that have helped me. Thanks to all the former and current students, postdocs, and staff in the CIGI Laboratory and CyberGIS Center. I have especially enjoyed working with Tao Lin and Yan Liu on many technical implementations in this thesis. I would also like to thank Susan Etter, Matthew Cohn, and Denise Jayne for their help in making my PhD journey smooth at UIUC.

This dissertation research is based in part upon work supported by the U.S. National Science Foundation under grant numbers: 0846655, 1047916, 1429699, and 1443080. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Last but most certainly not least, I thank my family for their unending love and support during this long, challenging, and rewarding journey. Special thanks to my wife Mingjing and our daughter Heather. I feel so lucky to have both of you and I never tell you enough how much I love you. Finally, I owe many thanks to my parents and Mingjing's parents. Thank you for being the most amazing parents anyone could ever ask for.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION AND BACKGROUND

1.1 MOTIVATION

Spatial decision-making is often influenced by complex geographic contexts, and has made extensive progress on leveraging geographic information science and systems (GIS) through the synthesis of geospatial data and analysis, domain-specific knowledge and models, and advanced information and computing technologies (Densham 1991; Jankowski et al. 1997; Leung 2012). During the synthesis, uncertainties from data, models, and computation are inevitably introduced. There is an increasing need for taking into account uncertainty in spatial decision support. Yet, most spatial decision support tools and systems are developed without paying much attention to data uncertainty (Fisher 1999; Shi et al. 2003; Shi 2009) and their propagation through spatial models (Heuvelink et al. 1989; Aerts et al. 2003). A major focus of this thesis research is therefore to establish innovative spatial decision-support capabilities for identifying, quantifying, and representing such uncertainty.

Various sources of uncertainty exist in spatial data, models, and their interactions during spatial decision-making processes (Ascough et al. 2008; Uusitalo et al. 2015). Given the limited knowledge of future outcomes, spatial decision making often requires data from predictive models (e.g., statistical models, simulation models) to be integrated with decision models. However, such data (e.g., weather, agricultural yields, movement of people) representing complex and dynamic spatiotemporal process are subject to uncertainty, which often require statistical analysis such as hierarchical models to quantify. Uncertainty also exists in the formulation of spatial decision models. Many spatial decision-making problems are formulated as optimization models and solved by either exact or heuristics methods. For example, spatial optimization models have been

established to support decision making in land use allocation (Ligmann-Zielinska et al. 2008), emergency response (Church and Cova 2000), facility location and supply chain management (Melo et al. 2009; Lin et al. 2013). Some of the models are developed with consideration of stochastic features while others are purely deterministic. As uncertainties are involved in spatial optimization, deterministic models should be replaced by more effective approaches such as stochastic programming (Birge 2011) and robust optimization (Ben-Tal et al. 2009) to better represent the problems. Furthermore, when a various source of data uncertainty are interacting with spatial decision support models, uncertainty will propagate and eventually complicate the decision making process.

Developing a spatial decision support system with uncertainty quantification brings new requirements in data analytics, model development, and computation. As one of the major challenges, uncertainty quantification for complex spatial models often requires computationally intensive model simulations such as uncertainty and sensitivity analyses (Saltelli et al. 1999; Lilburne and Tarantola 2009). Moreover, when model-driven approaches such as stochastic programming based spatial optimization are developed to handle data uncertainty in spatial decision support, the spatiotemporal perspective of the uncertainty is not well addressed. Thus, spatiotemporal data analytics need to be involved as a major part of spatial decision support.

CyberGIS, defined as GIS based on advanced computing and cyberinfrastructure (Wang 2010; Wang 2013), provides a desirable framework to address the new requirements in spatial decision support systems by seamlessly integrating a highly interactive online GIS user environment, friendly cyberinfrastructure access, and data and computing intensive spatiotemporal analytic methods. Therefore, cyberGIS becomes a fundamental basis for this study to build on.

1.2 RESEARCH QUESTIONS

The primary objective of this PhD thesis research is to develop an uncertainty-aware spatial decision support framework for supply chain optimization. The cyberGIS framework and related method are evaluated based on a case study of biomass-to-bioenergy supply chain optimization at the county level in the US to resolve the synthesis challenge in multiple spatial decision support scenarios. In addition, spatiotemporal analysis of agricultural yield will be discussed and integrated into an existing biomass-to-bioenergy supply chain optimization problem. Through the development of this thesis, I hope to address the following research questions:

1. How does cyberGIS enable complex spatial decision making in the application of supply chain optimization, particularly in the case of biomass feedstock provision?
2. How does cyberGIS-enabled decision support systems make uncertainty and sensitivity analysis computationally efficient on large scale supply chain optimization problem and enable new discoveries?
3. How to develop an innovative method to optimize supply chain by accounting for spatiotemporal explicit uncertainties?

To better address the above research questions, this research is designed to achieve the following research objectives:

1. Design and develop a cyberGIS-enabled spatial decision support system – CyberGIS-BioScope – to facilitate strategic planning on large-scale biomass-to-bioenergy supply chain at county level in the US with multiple spatial decision support scenarios;
2. Investigate methodologies in uncertainty and sensitivity analysis, and leverage the developed cyberGIS-enabled spatial decision support system to efficiently understand the

propagation of uncertainty from model inputs to outputs and the corresponding sensitivity estimation;

3. Develop a Bayesian statistical model to quantify spatiotemporal uncertainty in the supply side and integrate the results with a stochastic programming based supply chain optimization problem.

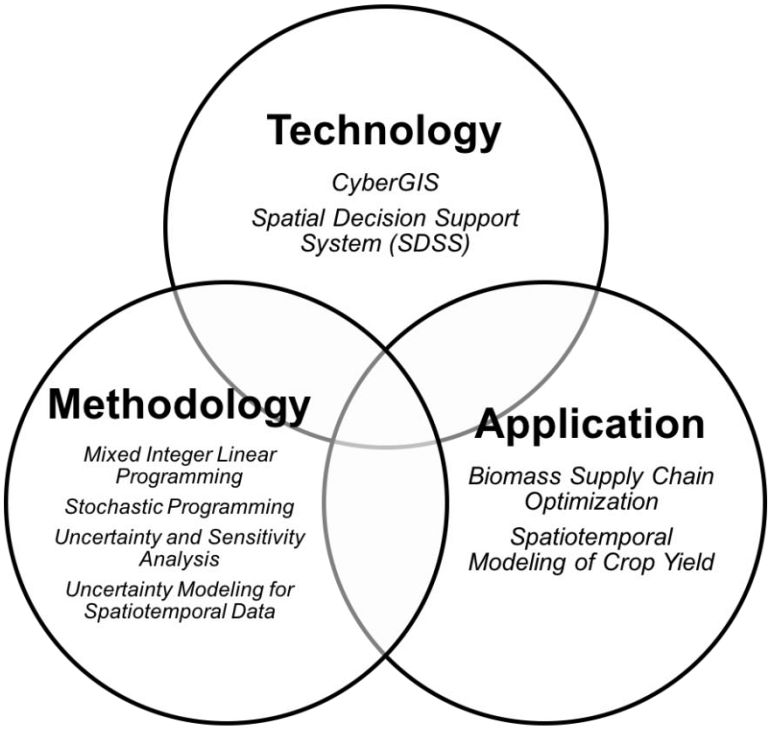


Figure 1.1: The integration diagram of this study

This research will establish a novel cyberGIS framework for resolving supply chain optimization with uncertainty quantification. The new architectural design of the proposed cyberGIS framework will provide guidelines for a new generation of spatial decision support systems which integrate capabilities of manipulating spatial data, enhancing result visualization and sharing, quantification of uncertainties and their propagations, as well as making complex decision models and advanced

computing resources easily accessible for decision makers. Furthermore, the theoretic contribution of considering spatiotemporal data uncertainty in stochastic programming will shed light on supply chain optimization problems that involve uncertainty, which moves beyond conventional simulation approach to uncertainty propagation. Although case studies are illustrated using examples in crop yield analysis and biomass-to-bioenergy supply chain management, the methodological framework is expected to have far-reaching effects beyond the domain of case studies.

1.3 BACKGROUND

1.3.1 Spatial Decision Problems and Models

Spatial decision making can be roughly defined as a specific type of decision making in which the decision implies the selection among several potential actions or alternatives that are associated with the geographic context, i.e., locations in space (Chakhar and Mousseau 2003). From individuals to organizations, people make spatial decisions such as routing vehicles, selecting a neighborhood to live in, choosing land development strategy, locating facilities and allocating resources, and managing infrastructure. Spatial decision problems are of great interest to researchers from diverse disciplines (e.g., agriculture engineers, economists, planners, ecologists, politicians, etc.) and different paradigms and modeling approaches have been established in each domain to solve real-world problems.

A large number of real-world spatial decision problems are abstract and structured as models using statistical methods, mathematical models, heuristic procedures, algorithms, and so on. Depending on the quantity and the type of available information, these problems can be classified into deterministic problems, stochastic problems and fuzzy problems (Pipkin 1991; Munda 2012;

Malczewski 1999), which are based on perfect, probabilistic and imprecise information respectively. As various spatial decision models are under active development, representing spatial decision problems and models to analysts and decision makers is not an easy process. It requires a user-friendly and flexible environment to be developed based on the integration of cartographic visualization tools, spatial query tools, collaborative decision making capabilities, as well as analytical models. This single user-centric environment is often known as spatial decision support system.

1.3.2 Spatial Decision Support Systems and CyberGIS

Compared with traditional decision support systems, spatial decision support systems (SDSS) set new requirements for accessing, querying and displaying various types of geographic data over multiple scales, and engaging users with interactive visual analytics (Chakhar and Mousseau 2003). Geographic information system (GIS), with its ability to create, manipulate, analyze, store and display various types of geographic information, has been widely incorporated into spatial decision-making since computer-based mapping technologies became powerful decision-making tools (Longley et al. 2005; Gewin, 2004). Taking advantage of spatial data management functionalities in GIS, spatial decision support systems (SDSSs) integrate various spatial decision support models, databases and assessment tools under a graphical user interface to support decision-makers with management actions (Matthies et al. 2007). Most existing SDSSs are primarily designed and developed based on single desktop or server mode, which limits the capabilities of handling massive and various geographic data, solving computationally intensive decision support models, and allowing collaborative decision making.

CyberGIS, denoted as cyberinfrastructure-based GIS, has emerged as a new generation of GIS harnessing advanced cyberinfrastructure resources, i.e., heterogeneous parallel and distributed computational resources, for solving computationally intensive and collaborative geospatial problems (Wang and Liu 2009; Wang 2010; Wang et al. 2013). Cyberinfrastructure is capable of providing huge computational powers, but it has not been developed to be used by researchers to solve their domain-specific problems in an easy manner. CyberGIS bridges this gap by connecting cyberinfrastructure to GIS and related spatial analysis and modeling (Wang and Liu 2009). Leveraging the capabilities of cyberGIS, the first part of this research will present a technological framework of implementing a cyberGIS-enabled SDSS where spatial data, service, visualization and computing resources are integrated to provide decision-makers with interactive and responsive decision-making experience. Challenges in handling spatial data and tools interoperability, computational complexity and user interactivity during the cyberGIS integration process will be discussed with technical details.

1.3.3 Spatial Optimization and Supply Chain Modeling

Spatial optimization has long been an important subdomain in the discipline of geography (Church 2001; Xiao 2009; Church and Murray 2009; Tong and Murray 2012). In the fields of transportation, location modeling, medical geography, land use planning, political geography and others, spatial optimization has been widely studied by researchers from within and outside the GIScience domains. Compared with general optimization problems in operations research, spatial optimization relies on optimization techniques to structure and solve problems where spatial context is crucial (Tong and Murray 2012). Supply chain optimization is among the categories of spatial optimization problems, with the goal to ensure the optimal operation of product

manufacturing and distribution by minimizing operating costs or maximizing production profits (Shapiro 2006). In supply chain optimization modeling, many decision variables and constraints are represented with spatial context (e.g., the location of facilities, transportation road network). Supply chain modeling has been widely applied to solve different kinds of business.

In this study, I would like to focus on two major challenges in the current and future development of supply chain modeling. The first one is the demand for web-based supply chain management systems that integrate diverse types of data, computationally intensive decision models, and collaborative problem-solving capabilities in a user-friendly environment. Another major challenge in uncertainty handling in supply chain modeling, which will be discussed in Section 1.2.5.

1.3.4 Uncertainty in Spatial Decision Making

Uncertainty is a complex term with many definitions and interpretations across knowledge domains and application contexts (Gahegan and Ehlers 2000). In a simple way, uncertainty is the lack of exact knowledge, regardless of what is the cause of this deficiency (Refsgaard et al. 2007). In literature, the term *uncertainty* is often ambiguously defined compared to related terms such as *data quality, reliability, precision, accuracy,* and *error*. Researchers tried to distinguish *error* from *uncertainty* based on whether the perfect information is known or not (Hunter and Goodchild 1993). However, in most cases, uncertainty is used as an umbrella term to describe all the related concepts (Zhang and Goodchild 2002). In many cases of spatial decision making, the involvement of uncertainty is desired. Without the consideration of uncertainty, the reliability of any decision becomes problematic (Matthies et al. 2007). Before modeling uncertainty, identifying the sources

of uncertainty is a prerequisite. In this study, the sources of uncertainty in spatial decision making are defined from two major perspectives – data and models.

Firstly, uncertainties exist in both spatial and non-spatial data. Data with measurement error is one example. Data processing, e.g., aggregation or interpolation or any algorithms introduced, will introduce uncertainty as well (Kwan 2016). Also, the way people perceive things might cause the data to be uncertain as well, for example, data related to experts' opinions. Furthermore, some decision models may require input data collected based on predictions, such as weather in next couple weeks, house price in the coming five years, or crop yield for the next growing season. Since our knowledge of future involves uncertainty, such data derived from predictive models are also subject to uncertainty.

Data uncertainty is often expressed in the form of probability distribution that indicates how likely each of the possible outcomes is (Uusitalo et al. 2015). When considering data uncertainty with deterministic spatial analysis and decision models, considerable research efforts have been made on the quantification of uncertainty propagation using uncertainty and sensitivity analyses. Examples of application domains include flood forecasting (Crosetto and Tarantola 2001), groundwater contaminant modeling (Lilburne and Tarantola 2009), land suitability evaluation (Ligmann-Zielinska and Jankowski 2014), and hazardous waste disposal planning (Gómez-Delgado and Tarantola 2006). In these literature, uncertainty and sensitivity analyses are employed as an integrated approach where uncertainty analysis aims at understanding the variability of outcomes given model input uncertainty, while sensitivity analysis focuses on quantifying the impact of each input that are responsible for this variability. However, few previous pieces of research have incorporated the uncertainty and sensitivity analyses into an SDSS. In this study, the integration part is one of the major contributions. Specifically, the integrated system takes

advantage of advanced cyberinfrastructure to solve the major computational issue (i.e., large number of Monte Carlo simulations) encountered in the uncertainty and sensitivity analyses, which could be critical for modeling uncertainty propagation in complex spatial decision models.

Secondly, uncertainties exist in decision models since models are always abstractions of the natural system. Often times, it is difficult to have perfect knowledge about the complete processes of the model and its parameters. Uncertainty in model parameters can be accounted for in a similar way as data uncertainty using probabilistic representations, while uncertainty about the model structure is complicated and it requires stochastic models to be developed instead of deterministic ones (Chatfield 2006). Developing stochastic supply chain optimization models will be discussed with details in the next subsection.

1.3.5 Supply Chain Optimization under Uncertainty

Uncertainty and sensitivity analyses is an effective approach to quantifying uncertainty propagation and derive sensitivity of each uncertainty sources with respect to the results. However, in supply chain optimization, uncertainty and sensitivity analyses do not provide direct solutions for decision-maker. Instead, distributions of decision variables and outcomes are presented with the sensitivity of each uncertain factors quantified. To derive optimal solution in the presence of uncertainty, stochastic models, e.g., stochastic programming (Santoso et al. 2005), robust optimization (Pishvaee et al. 2011) or multi-objective optimization with risk function (Azaron et al. 2008), are commonly used to tackle supply chain optimization problems. These models are designed with capabilities of capturing uncertainty information in the model construction instead of treating the model as a black box in uncertainty propagation and sensitivity analysis.

In stochastic programming models, decisions are considered at multiple stages. As an example of a two-stage stochastic programming problem, decision maker takes some actions in the first stage, after which uncertain factors start affecting the outcome of the first-stage decision. A recourse decision can then be made in the second stage that compensates for any bad effects that might have been experienced as a result of the first-stage decision (Birge and Louveaux 2011). In supply chain optimization, previous studies (Barbarosoğlu and Arda 2004; Santoso et al. 2005; Sodhi and Tang 2009; Marufuzzaman et al. 2014) have developed variations of two-stage stochastic programming models, in which facility locations are considered as long-term decisions that need to be fixed at the first stage, and supply chain logistics are considered as recourse decisions that can be adjusted to cope with uncertain situation.

Although a considerable number of stochastic models have been developed in literature to account for uncertainty, models that explicitly account for uncertainty in spatiotemporal data are limited. Spatiotemporal data collected from the agricultural process or ecological phenomenon are subject to uncertainty and they need to be modeled along with the underlying spatiotemporal process affected by many environmental variables. Such complexities often require statistical analysis such as hierarchical models (both Bayesian and non-Bayesian) to quantify (Cressie et al. 2009). Therefore, the last piece of this study is trying to develop an effective approach to incorporating statistical analysis of spatiotemporal data into stochastic supply chain optimization models.

1.4 ORGANIZATION

This dissertation consists of three papers addressing how cyberGIS facilitates decision support and enables new discoveries in the case study of biomass supply chain optimization with the

consideration of uncertainty. First, a cyberGIS-enabled spatial decision support system is designed and developed to provide interactive and collaborative decision support in biomass provision at the county level in the United States (Chapter 2). Second, through harnessing the computational power of cyberGIS, uncertainty and sensitivity analysis are conducted on the supply chain of a potential bioenergy crop, Miscanthus, in Illinois to efficiently understand how different sources of uncertainty impact on optimal supply chain decisions (Chapter 3). Third, to consider spatiotemporal explicit uncertainty in the supply side, a Bayesian statistical model is developed and the results of the spatiotemporal analysis are integrated with a stochastic programming method to optimize corn stover supply chain in four Corn Belt states, i.e. Illinois, Iowa, Missouri, and Indiana, in the Midwestern US (Chapter 4). Finally, major findings and contributions of this research are discussed (Chapter 5).

Chapter 2 titled "A CyberGIS-enabled Spatial Decision Support System for Biomass Supply Chain Optimization" demonstrates an interactive and collaborative system for supply chain optimization by addressing challenges exist in 1) the limited interoperability between bioenergy models and geographic information systems (GIS); 2) interactive scenario construction, evaluation and sharing; and 3) complex optimization problem solving that requires advanced cyberinfrastructure resources to support interactive decision making. The developed system, CyberGIS-BioScope, takes advantage of cyberGIS capabilities to process and analyze spatial data and enhance visualization and sharing of optimization results. The integrated environment makes the complex optimization model and advanced cyberinfrastructure resources easily accessible for agricultural scientists and decision makers, and thus accelerates their scientific discovery and decision-making processes.

Chapter 3 titled “A CyberGIS Approach to Uncertainty and Sensitivity Analysis In Biomass Supply Chain Optimization” describes a cyberGIS approach to optimize biomass supply chains under uncertainties. This approach has been implemented as a decision support system through integration of data management, mathematical modeling, uncertainty and sensitivity analysis, scenario analysis, and result representation and visualization. An optimization modeling analysis of 7,000 scenarios using Monte Carlo methods has been conducted to quantify the uncertainty and sensitivity impact of various input factors on ethanol production costs and optimal biomass supply chain configurations in Illinois, United States. Leveraging high performance computing power through cutting-edge cyberGIS software, what-if scenario analysis has been evaluated to make decisions in case of unexpected events occurring in the supply chain operations.

Chapter 4 titled “Biomass Supply Chain Optimization with Uncertain Biomass Availability” presents an innovative approach that captures spatiotemporal data uncertainty in a spatial supply chain optimization problem. Based on a Bayesian hierarchical model that accounts for spatial and temporal random effects in the coefficients when conducting regression analysis, this novel approach captures spatiotemporal explicit uncertainty associated within the crop yield estimation and leverage such information to construct different supply availability scenarios when formulating the stochastic programming optimization. The proposed method is applied to a biomass supply chain problem that aims to optimize supply chain infrastructure configurations, biomass feedstock provision and logistics operations using corn stover as the bioenergy crop. Variation in corn yield, farmer participation rate, and collectable corn stover rate are considered as key uncertain factors for biomass feedstock supply.

Chapter 5 concludes the dissertation by synthesizing the major findings of this research and potential directions for future research opportunities.

CHAPTER 2

A CYBERGIS-ENABLED SPATIAL DECISION SUPPORT SYSTEM FOR BIOMASS SUPPLY CHAIN OPTIMIZATION[1]

This chapter describes the development of a cyberGIS-enabled spatial decision support system for optimizing biomass feedstock provision. The developed system, CyberGIS-BioScope, aims to facilitate complex problem solving in biomass supply chain through an interactive and collaborative decision-making fashion. CyberGIS-BioScope takes advantage of cyberGIS capabilities to process and analyze spatial data and enhance visualization and sharing of optimization results. Meanwhile, the integrated environment makes the complex optimization model and advanced cyberinfrastructure resources easily accessible for agricultural scientists and decision makers, and thus accelerates their scientific discovery and decision-making processes. This implementation example could be served as a protocol for further integration development of cyberinfrastructure, operations research, and geospatial analysis and modeling.

This chapter cannot be completed without successful teamwork. The team members are Hao Hu, Tao Lin, Yan Liu, Luis Rodríguez, and Shaowen Wang. Mr. Hu led the overall research design, development of the system, and manuscript writing. Dr. Lin contributed to the original source code of the BioScope model that serves as the foundation for biomass supply chain optimization. Mr. Liu led the service integration and assisted the draft revision. Drs. Lin, Rodríguez, and Wang

[1] Reprint, with permission, from Hu et al., 2015, "CyberGIS-BioScope: a cyberinfrastructure-based spatial decision-making environment for biomass-to-biofuel supply chain optimization", Concurrency and Computation: Practice and Experience 27 (16), 4437-4450

participated in the overall research design and results discussions, and led the draft revision. The content of this chapter has been published in a journal paper (Hu et al., 2015).

Abstract. *Biomass, e.g. energy crops, forests and agricultural residues, has emerged as a renewable energy option to alleviate the consumption of limited fossil fuel resources and the consequent environmental issues. Designing an effective and efficient biomass-to-biofuel supply chain involves sophisticated decision-making processes, often requiring collaborative work on data integration, model specification, scenario analysis, and coordinated implementation and management. To establish an integrated system for such work, challenges exist in 1) the limited interoperability between bioenergy models and geographic information systems (GIS); 2) interactive scenario construction, evaluation and sharing; and 3) complex optimization problem solving that requires advanced cyberinfrastructure resources to support interactive decision making. To resolve these challenges, this paper describes CyberGIS-BioScope: an interactive and collaborative cyberGIS-based spatial decision-making environment for biomass-to-biofuel supply chain optimization. CyberGIS-BioScope takes advantage of cyberGIS capabilities to process and analyze spatial data and enhance visualization and sharing of optimization results. Meanwhile, the integrated environment makes the complex optimization model and advanced cyberinfrastructure resources easily accessible for agricultural scientists and decision makers, and thus accelerates their scientific discovery and decision-making processes.*

Keywords. CyberGIS; biomass supply chain; spatial decision making; web-based gateway environment

2.1 INTRODUCTION

Biofuel has gained attention given its potential contribution to energy independence and rural economic development. Increased demand for biofuel, however, requires a new system infrastructure design for a biomass-to-biofuel supply chain. To ensure efficient and effective biofuel production, considerable research efforts have been made on biomass-to-biofuel supply chain optimization to identify best solutions for biomass feedstock provision, infrastructure investment, and logistics operation (Marvin et al., 2012; Nagel, 2000; Brechbill et al., 2011; Huang et al., 2010; You and Wang, 2011; Kim et al., 2011; Lin et al., 2013).

To support decision-making on problems such as the biomass supply chain, decision support systems (DSS) are developed to integrate various models, databases and assessment tools under a graphical user interface. When integrating functionalities with geographic information systems, DSSs become spatial decision support systems (SDSSs) (Matthies et al., 2007). SDSSs have been established to facilitate biomass resource management (Zambelli et al., 2012) and multi-criteria infrastructure planning (Perimenis et al., 2011). However, most existing decision support applications for biofuel development have not been well integrated with supply chain optimization modeling because of the following major challenges:

Data interoperability. Geospatial and engineering data are distributed from various sources. Agricultural data collected from spatial data services may not be ready for directly feeding into decision-making models until proper preprocessing steps are taken. Meanwhile, output data from models need to be standardized for visualization and sharing. Therefore, interoperability between agricultural engineering data and spatial visualization and analytic tools becomes crucial to enable the evaluation and sharing of biomass-to-biofuel supply chain results.

Computation complexity. Biomass-to-biofuel supply chain design is often formulated as an optimization problem that desires to be solved efficiently in the context of responsive decision support. Most existing biomass supply chain decision systems are single machine based without dedicated computing resources to either reduce the single job execution time or support large number of concurrent job submissions.

User Interactivity. A user-centered decision-making environment requires not only a rich user interface with complex formatting through multimedia, but also a high level of interaction that enhances user involvement. To our knowledge, none of the previous studies has developed a highly interactive web-based user environment for biomass-to-biofuel supply chain decision-making that supports step-by-step scenario construction and what-if scenarios analyses in a geodesign (Steinitz, 2012) fashion.

Leveraging advanced cyberinfrastructure (CI), cyberGIS (Wang, 2010; Wang et al., 2013) provides a desirable framework to tackle the aforementioned challenges by synergistically integrating a highly interactive online GIS user environment, seamless cyberinfrastructure access, and data and computing intensive spatial analysis methods through three major modalities: CyberGIS Gateway, GISolve middleware, and CyberGIS Toolkit. Based on the cyberGIS framework, a science gateway application CyberGIS- BioScope is developed on top of BioScope model (Lin et al., 2013) to provide the benefit of scalable and interactive spatial decision support for biomass supply chain optimization in this study.

Advanced CI resources are needed for meeting the following computational requirements in spatial decision-making:

- The computation of an individual model is time consuming, requiring great computing power. BioScope optimization is a mixed integer linear programming (MILP) problem that

is known to be computationally intractable, namely, it cannot be solved in polynomial time based on the existing algorithms;

- Uncertainty analysis centered on the BioScope model may generate a large number (e.g., hundreds or even thousands) of individual model runs , each requiring invocation of a series of spatial data services and computations of the BioScope optimization model; and
- Spatial decision-making is often an interactive process between decision makers and model computation, requiring responsive computation and presentation of model results (e.g., in minutes instead of hours or days).

The goal of CyberGIS-BioScope application development is to establish a cyberGIS-empowered spatial decision-making environment via geodesign that provides the following major capabilities:

- An interactive and collaborative web environment that combines the specification and integration of geospatial and engineering data sources, parameters, and models to facilitate flexible and easy-to-use construction of scenario analysis by a group of cross-disciplinary decision-makers;
- A diverse set of result visualization representations including map layers, result tables and charts for a comprehensive understanding of biomass supply chain optimization output;
- A seamless computation management framework that can leverage advanced CI resources for intensive computation generated by community use, batch scenario studies, and sophisticated analyses such as uncertainty and sensitivity analyses that often involve Monte-Carlo simulation of hundreds or even thousands of individual model runs; and

- A computational intensity profile that serves as a guideline in cyberinfrastructure resource allocation in order to request appropriate amount of computing power and maximize resource utility for solving the biomass-to-biofuel supply chain optimization problem.

Figure 2.1. A Three-stage biomass supply chain for biofuel production.

2.2 BACKGROUND

BioScope (Lin et al., 2013), as one of the state-of-art biomass supply chain optimization models, illustrates a typical three-stage biomass supply chain with two logistics phases involved (Figure 2.1). In the first logistics phase, raw biomass feedstock are supplied from distributed farms to centralized storage and preprocessing (CSP) facilities after harvesting. As an intermediate stage of the supply chain, CSP facilities are designed to reduce transportation cost through preprocessing raw biomass to provide ground biomass with tapping. Subsequently in the second logistics phase, preprocessed biomass is delivered to biorefineries where biomass is converted into biofuel products. The BioScope model aims to optimize the location and amount of biomass supply, the location and capacity of storage and biorefinery facilities, and transportation flow between each stage with the consideration of constraints in supply availability, capital and operational cost, conversion technologies, and biofuel demand. The BioScope model is formulated as a mixed integer linear programming (MILP) problem with a linear objective function and a list of linear constraints in which integer and contiguous decision variables need to be solved. Specifically, the

objective function is a linear combination of biomass purchase costs, transportation related costs, CSP site related costs, and biorefinery related costs. Decisions such as biomass capacity for a facility or transportation flows (e.g., amount of biomass delivered from farm i to facility j) are modeled as continuous variables while facility location decisions (e.g., should facility at site j be constructed) are restricted to be integers. Details of the Bioscope model formulation can be found in Lin et al. (2013).

Given the advantages in spatial data handling and visualization, GIS has been integrated in the development of spatial decision support system in bioenergy research. For example, Voivontas et al. (1998) developed a GIS-based decision support system to estimate the power production potential of agricultural residues. Freppaz et al. (2004) and Frambo et al. (2009) integrated GIS with mathematical programming methods to investigate forest biomass exploitation for energy production. However, GIS mostly behaves as an external tool in spatial data aggregation, road network distance measurement and spatial data visualization in bioenergy researches (Marvin et al., 2012; Lin et al., 2013; Wang et al., 2013). Many developed decision support softwares integrate decision models with map-based interface based on desktop-based ArcGIS software (ESRI), which limits the accessibility for collaborative use (Frombo et al., 2009; Tibaa et al., 2010). Furthermore, computational performance becomes a requirement when the underlying decision model is complex to solve. A majority of existing biomass supply chain optimization problems are formulated as mathematical programming models such as MILP. Since MILP problem is computationally intractable (i.e. NP-hard), computational efficiency becomes a bottleneck when the problem structure is complex. As yet, few efforts have considered leveraging advanced cyberinfrastructure to solve biomass supply chain optimization problems. CyberGIS provides an

alternative by not only bringing GIS functionalities for spatial data handling and visualization, but also providing access to advanced cyberinfrastructure with scalable computation resources.

MILP problems are commonly solved using linear programming techniques (Williams, 2009) (e.g., branch-and-bound and branch-and-cut to tighten the linear programming relaxation) where the original MILP problem (root) is relaxed to sub-MILP problems (search nodes) based on selected branching variables recursively until an integer solution (leaf) is found. The fact that each search node can be solved independently provides potential opportunity for implementing parallelism of solving MILP problem on multiple cores or machines. CPLEX and Gurobi are two leading commercial optimization solvers and they both provide free license for academic use. The latter one is selected as the optimization solver in the CyberGIS-BioScope application. With the latest release of Gurobi 6.0, parallel problem solving is supported in a distributed framework where multiple machines work together to solve a single MILP problem.

2.3 APPLICATION DEVELOPMENT

The development of CyberGIS-BioScope application follows a streamlined CyberGIS gateway application development framework and integration process. This framework includes common software engineering practice for builds and tests, cyberinfrastructure-based portability tests across a variety of operating systems and software libraries, scalability tests, gateway application development, registration and deployment, application service publishing and CyberGIS Gateway integration (Wang et al., 2013; Liu et al., 2015). The unique requirements in CyberGIS-BioScope leads to the development of a rich-client based spatial data visualization module and a job scheduler-based computation workflow tool for complex scenario analysis. The architecture of

CyberGIS-BioScope has three tiers: CyberGIS Gateway, cyberGIS services, and the cyberinfrastructure-based computation (Figure 2.2).

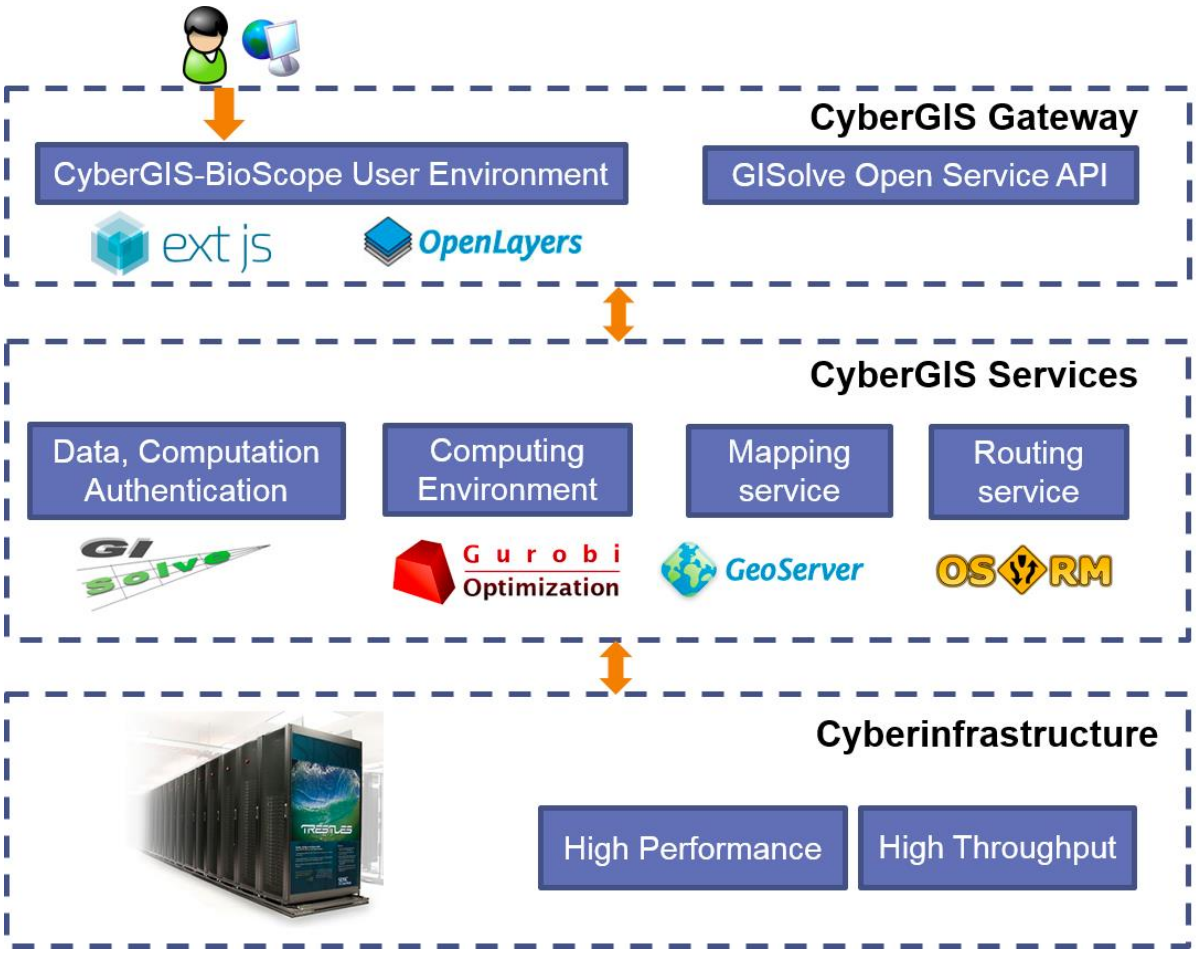


Figure 2.2. CyberGIS-BioScope Application Architecture.

2.3.1. User Environment

The user interface of CyberGIS-BioScope was developed based on HTML5 technologies, i.e., Sencha ExtJS (http://sencha.com), a JavaScript web platform development framework, and OpenLayers (http://openlayers.org), an open source JavaScript library for map data displaying and handling in web browser. The user environment development adopts the client-side model-view-control (MVC) model, which greatly improves the user interaction capabilities within the web browser. The user environment interacts with standard web services (e.g., REST and Open

Geospatial Consortium (OGC) Web Mapping Service (WMS), Web Feature Service (WFS), and Web Coverage Service (WCS)) on the backend using Hypertext Transfer Protocol (HTTP) and asynchronous JavaScript and XML (AJAX). The user environment consists of two major components: 1) the BioScope scenario construction, and 2) visualization of optimal biomass-to-biofuel supply chain solutions.

To construct a scenario analysis, a user needs to go through four major steps to identify input data and specify model parameters: 1) select candidate biomass supply counties; 2) select candidate centralized storage and preprocessing (CSP) facilities; 3) select candidate biorefinery facilities; and 4) specify other non-spatial parameters such as cropland usage rate (i.e., percentage of cropland converted to growing bioenergy crops) and unit transportation cost. The user interface presents both spatial and data view of model inputs. The first three steps require a set of map-based spatial feature selection actions, including map panning, zooming, box selection, individual spatial object selection/deselection. Geospatial and engineering data attributes associated with selected spatial objects are then loaded dynamically by sending spatial queries to a published WFS that has biomass- related attribute data (e.g., bioenergy crop yield, procurement cost, cropland area at county level). An attribute table (left panel in Figure 2.3a) displays these attributes information of the selected region. Non-spatial data inputs in step 4 are fetched from an ExtJS form. Once all the required data and parameters are provided, a new scenario analysis job is created and submitted using GISolve Open Service API. Serving as the middleware between cyberGIS user environment and cyberinfrastructure-based resources and services, GISolve Open Service APIs are used to provide a streamlined integration process for application registration and configuration, code invocation and scheduling, as well as data transfer and visualization. Details on using the GISolve

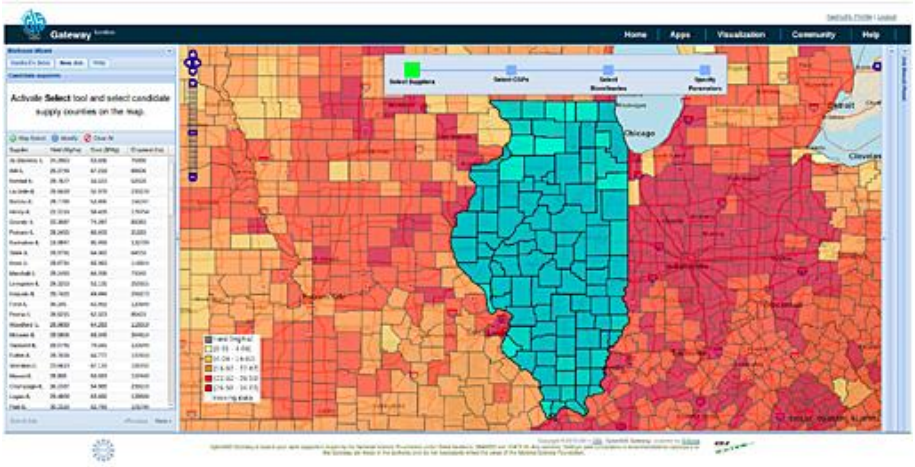
Open Service APIs as a streamlined service integration to cyberinfrastructure can be found in Wang et al. (2013).

Currently, CyberGIS-BioScope has incorporated county-level data of Miscanthus yield (Miguez et al., 2012), cropland area (http://quickstats.nass.usda.gov), and Miscanthus production costs across the contiguous U.S., including 48 states that are composed of 3,109 counties, where users can select from to perform Miscanthus supply chain case studies. Moreover, users may alternatively provide their own data through editing the attribute table for the selected region of interest. Road network data from OpenStreetMap (http://openstreetmap.org) are integrated to measure transportation distance and enable the visualization of supply chain logistics.

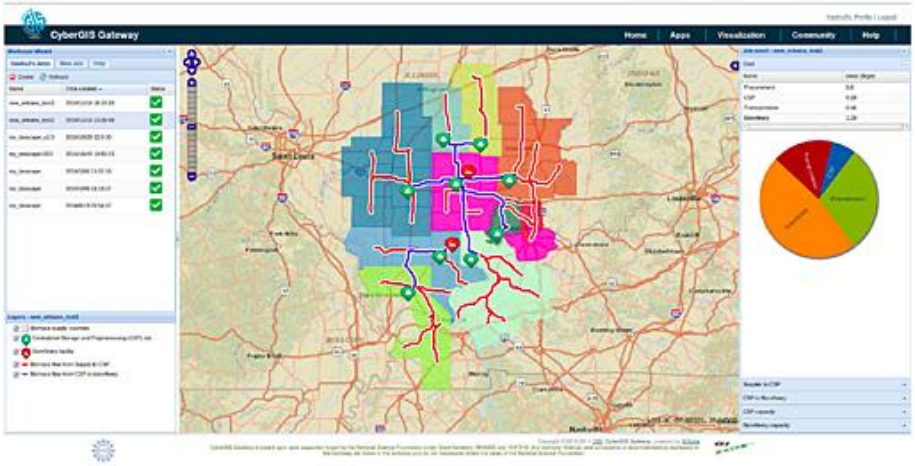
Top-left panel in Figure 2.3b shows a list of jobs and information including the job name, created time and job status. User can switch among finished jobs to compare different biomass-to-biofuel supply chain scenarios. For a specific scenario, the map panel shows five result data layers: the optimal locations of biorefinery facilities (red icons), the optimal locations of CSP facilities (green icons), biomass transportation flows from supply counties to CSP (red path), biomass transportation flows from CSP to biorefinery (blue path), and biomass supply counties (counties with the same color share one CSP facility). Breakdowns of biofuel production cost, capacity of biofuel facilities and transportation flow information are presented to users as a dynamic pie chart and data tables (Figure 2.3b).

In biomass-to-biofuel supply chain decision-making, some input variables are uncertain and subject to change given spatial and temporal variations. Changes in biomass yield, procurement price, market demand, transportation cost, and processing technology could significantly impact the overall production cost and spatial configuration of biomass supply and processing networks, which further complicates the assessment of decisions. Often user has to perform comparisons to

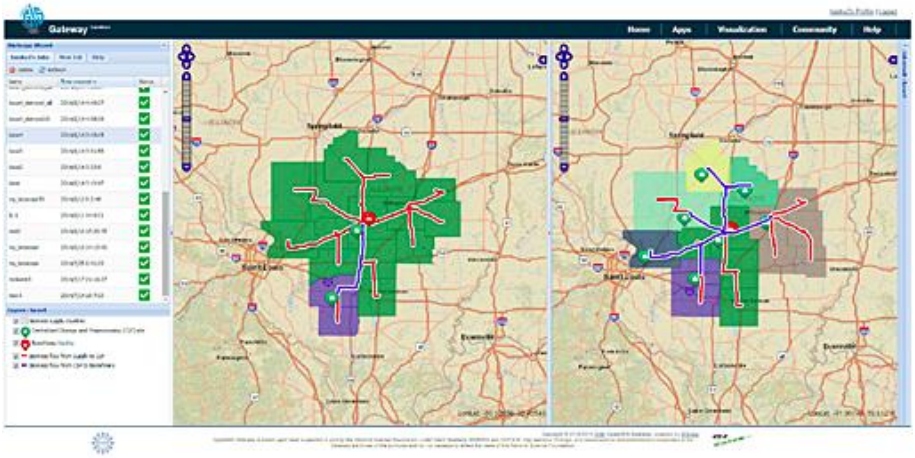
understand how the optimal supply chain pattern changes accordingly. For instance, by increasing the unit transportation cost to a foreseeable price predicted by the market, the model tends to select more CSP facilities close to the biomass supply sites in order to balance the cost spent in transportation. It then becomes interesting to learn the location of additional CSP facilities and how the biomass logistics changes subsequently. To intuitively visualize such analyses performed by users, a multiple map panel feature (Figure 3b) is developed through displaying two map panels that are synchronized in spatial scales and layer settings but loading results from two different scenarios.



(a)



(b)



(c)

Figure 2.3. CyberGIS-BioScope GUIs: (a) Input data identification. Candidate counties are highlighted in light blue and attribute information is displayed in the left panel; (b) Biomass-to-biofuel supply chain result visualization; (c) Interface for two scenarios comparison analysis.

Different from conventional server-side WMS-based map layer rendering, the visualization of BioScope output requires various formats of result presentation, including map layers, location attributes, charts and diagrams to facilitate a comprehensive evaluation of one or multiple supply chain solutions. Retrieving these output attributes dynamically using the conventional WFS or WCS would introduce frequent client-server communications and uncertain delays. Furthermore, WMS gridizes a data layer as map image tiles for rendering and results in lower image quality than directly drawing output spatial objects using HTML5. In order to improve user experience, BioScope output data is stored in JSON and XML format and sent directly to the browser to present map layers, result tables, text content, and charts.

As a result, the developed CyberGIS-BioScope user environment is equivalently powerful as conventional desktop graphical user interface (GUI) in terms of user interaction capabilities, attributing to the rich-client web interface technologies adopted in CyberGIS Gateway.

2.3.2. Open Service Integration Framework

The BioScope model solving requires an integration of a series of geospatial data and processing services and cyberinfrastructure services. Input data to BioScope model are either fetched from or dynamically computed by distributed geospatial data and mapping service instances. Standard web services, i.e., HTTP, REST, and OGC web services, are established to allow an interoperable integration of input data generation and output data hosting, as well as visualization. These services are hosted in a private cloud at the CyberGIS Center at the University of Illinois and can be dynamically allocated on demand based on the user requests.

Specifically, the following web services are established to support an interoperable coupling of BioScope input and output data access and processing:

Routing service. Open Source Routing Machine (OSRM) (Luxen and Vetter, 2011), an open source high-performance routing engine, is deployed to find the shortest path and its geometry information for a given pair of coordinates by using OpenStreetMap road network data. It utilizes OpenMP to improve the performance for route calculation.

Geospatial and engineering data and mapping service. County-level crop yield, biomass procurement cost, and cropland area data for the continental U.S. are compiled and stored on a data server as vector datasets. GeoServer (http://geoserver.org), an open source mapping server, is deployed to present these datasets as OGC WMS and WFS. When a user selects certain spatial scale for a case scenario, these data are rendered as WMS data layers. Attribute information associated with the layers can be retrieved using separate AJAX calls as WFS requests. Data can also be downloaded using various geospatial vector file formats. This service has been deployed in the visualization virtual machine farm at the CyberGIS Center. It is noteworthy that although many biomass supply chain studies incorporated GIS technologies to enable map-based visualizations (Lin et al., 2013; Frombo et al., 2009; Noon and Daly, 1996), no similar efforts have been made to automate the process as presented in this study.

2.3.2.1. Cyberinfrastructure-based Computation.

The CyberGIS-BioScope software consists of several data and computing components and has been deployed on Trestles (http://trestles.sdsc.edu, retired) from XSEDE with high performance and high throughput computing capabilities. The computation workflow of BioScope is shown in Figure 2.4. Major spatial and non-spatial data required by the BioScope model are illustrated with their connection to different computing processes. Farm-level sampling tool, transportation

distance matrix generation, BioScope model solving, and statistical analysis of optimization results are identified as four major computing components:

Farm-level sampling. The BioScope model provides strategic decision support on the biomass supply chain configuration at county level, where a large number of distributed farms are aggregated into a single point for representing each county. For example, there are more than one thousand farms in Champaign County, Illinois in 2012. The data aggregation procedure results in the loss of locational information of farms and possible bias when calculating the transportation distance. Thus, an additional step in farm-level sampling is considered to generate location data of farms given the auxiliary farm-level information (i.e., farm numbers and cropland data layer from USDA National Agricultural Statistics Service, http://quickstats.nass.usda.gov/) within each county. Considering the uncertainty of biomass feedstock provision locations at farmland scale, random points that represent possible farm locations are sampled within the agricultural land. The farm-level sampling tool provides spatial support at finer scale for measuring the transportation distance between farms and biomass processing facilities. It is only required when conducting uncertainty analysis.

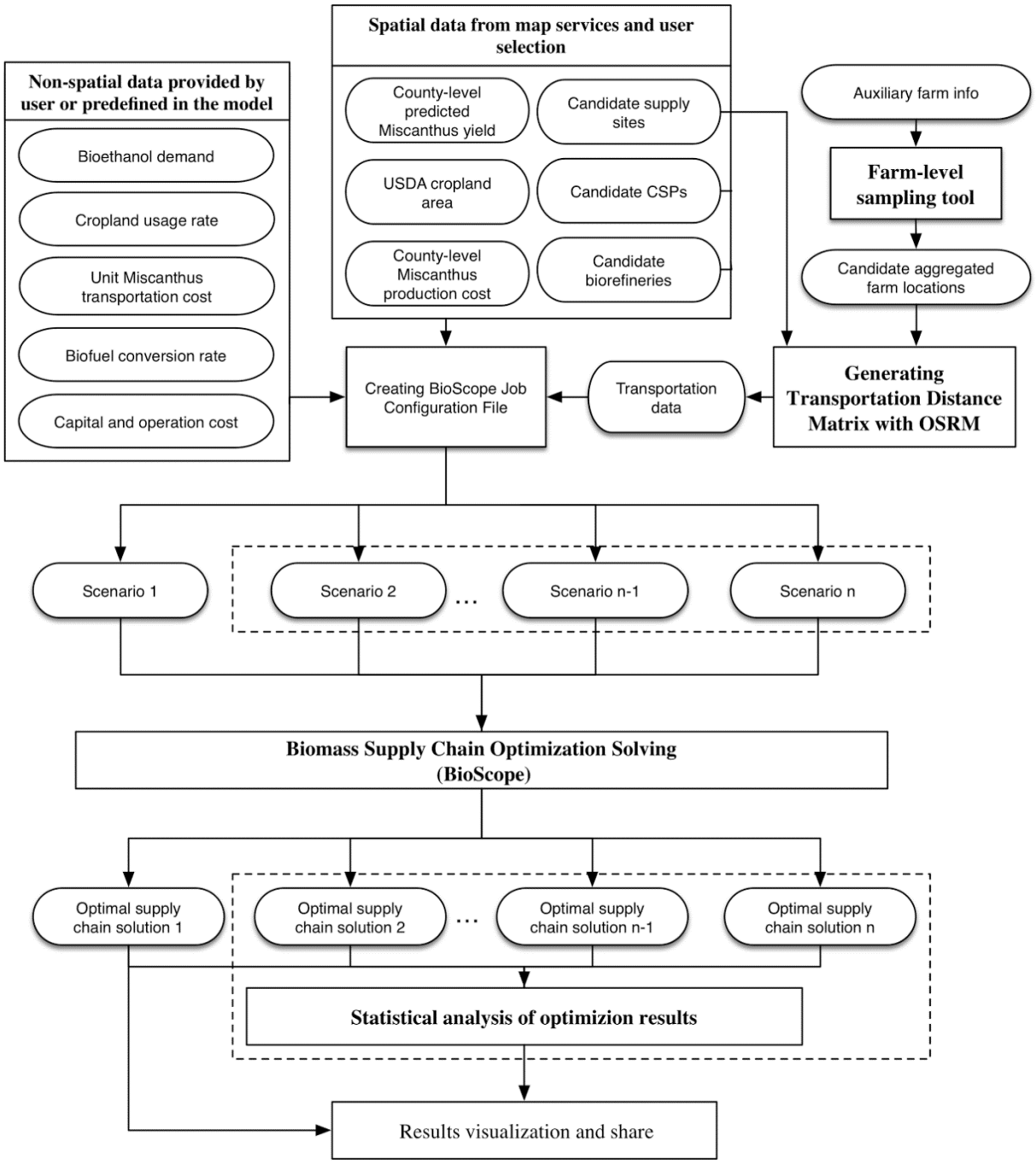


Figure 2.4. Computation workflow of CyerGIS-BioScope using Miscanthus as an application.

Generating distance matrix. The CyberGIS-BioScope measures the real road network distance when estimating the transportation cost. Based on the candidate supply sites, CSPs and

biorefineries locations defined by the user, a transportation distance matrix is dynamically generated through sending requests to the locally deployed OSRM routing service. Meantime, the route geometry information are retrieved from the request and saved to construct transportation layers when optimal logistics operation results are available.

Optimization problem solving. The solver for BioScope model is a commercial optimization software Gurobi, which has built-in parallelism to solve MILP problem. A typical model application for Illinois case study has a total of 12,961 constraint equations and 22,956 decision variables, including 816 binary variables. A typical biomass-to-biofuel supply chain scenario within Illinois that satisfies the requirement of 240 million gallons biofuel production takes 176 seconds using Gurobi solver (with optimality gap=0.01%) on an Intel(R) Core(TM) i7-4770 CPU @ 3.40GHz desktop machine. However, tuning the parameters (e.g., cropland usage rate or bioethanol demand) will increase the job execution time up to 2,527 seconds.

Statistical analysis tool. When multiple scenarios are constructed to consider model input parameters uncertainty, statistical analysis is required to quantify the impact of uncertain parameter(s) on the BioScope model results. The statistical analysis tool aims to provide statistical bounds on objective values (e.g., optimized annual biomass-ethanol cost) and decision variables (e.g., number of biomass processing facilities), and spatial configuration of biomass supply chain (e.g., the likelihood of each biomass processing facilities being selected). The invocation of statistical analysis in a BioScope model run is optional and only required when uncertainty analysis is specified.

2.3.2.2. Workflow Tool for Uncertainty Analysis

The execution of uncertainty analysis is implemented as a light-weight computation workflow by leveraging the job dependency features of job schedulers such as SLURM (http://slurm.schedmd.com) and PBS (http://openpbs.org) (i.e., sbatch (with option *-d*) and qsub (with option *-W* depend), respectively). For example, using sbatch, job B can be dependent on job A under three conditions: *afterok*, *afternotok*, and *afterany*: job B is executed after job A is finished with success, failure, or either, respectively. *afterok* is used to execute BioScope models because the previous step in transportation network generation must be successfully executed. The last step of uncertainty analysis can tolerate failures to a degree because the following step in statistical analysis does not require all the results from the sampled optimal solutions. *afterany* is used to represent the job dependency for this step. The generated job submission script, when executed, automatically substitutes job identifiers returned by job scheduler on the dependency list of subsequent jobs. This tool has been tested on a local cluster with six nodes (48 cores) at the CyberGIS Center.

2.3.3. Gateway Integration

The CyberGIS-BioScope application has been deployed in the production CyberGIS Gateway (http://gateway.cybergis.org) for community evaluation. At the Second International Conference of CyberGIS and Geodesign (August 18-21, 2014, Redlands, CA), CyberGIS-BioScope was demonstrated as a geodesign use case and received positive feedback from geographic information science community. As one of the CyberGIS Gateway applications, CyberGIS-BioScope takes advantage of the streamlined integrating process within the cyberGIS framework. On the UI side, an open mashup library framework has been implemented to support the development of user

interface components that can be customized for cyberGIS use. These components (e.g., job management panel, map visualization panel) can be shared across different cyberGIS applications, which significantly facilitate the implementation of the CyberGIS-BioScope application. The streamlined CyberGIS Gateway application development framework significantly simplifies the integration process. However, building a spatial decision-making system requires highly interactive user interface components to be developed within CyberGIS Gateway. Improving the productivity of developers is an evolving effort of the CyberGIS Gateway development environment.

The CyberGIS-BioScope application introduced two major requirements to the CyberGIS Gateway application development: 1) a tight coupling of external web services and computing workflow, and 2) the development of new visualization components based on raw geospatial output, instead of pre-rendered tiled data layers. A set of geospatial web services, such as the transportation routing service (i.e., OSRM), has been established as gateway web services and can be invoked within the computation workflow at runtime. New visualization components have been developed to load XML and JSON output of BioScope model to enrich the visualization of data layers on map and other visualization components (e.g., charts and tables) using a single data source. Compared to result visualization in other CyberGIS Gateway applications, CyberGIS-BioScope's direct result rendering using JSON greatly improved the rendering quality. This capability is being ported to a generic module in CyberGIS Gateway to provide a unified map creation and sharing framework to support other cyberGIS analysis results.

2.4 COMPUTATIONAL PERFORMANCE

In the gateway application integration, the scalability and computational performance testing is an important step to establish a computational intensity profile for GISolve middleware. This profile serves as a guideline for cyberinfrastructure resource allocation in order to request appropriate amount of computing power and maximize resource utility. As a spatial decision-making environment, CyberGIS-BioScope needs to return computation results in a specified time limit based on the responsiveness requirements from decision makers. The whole processing time of a CyberGIS-BioScope analysis can be decomposed into data preparation, the BioScope model execution, results processing and visualization, and the communication between each data and services. As the most time-consuming part, the BioScope model execution time largely depends on both the number of allocated processors and model parameters values specified in gateway user environment. Parameter values determine the solution space landscape and the computational efforts needed to search the solution space. Therefore, establishing the relationship between the model parameters and the number of processors for the aforementioned computational intensity profile is necessary for scheduling model computation in different decision-making scenarios. For example, the user environment components of CyberGIS-BioScope can provide a reasonable value range for each parameter under consideration based on the profile so as to make sure results can be returned within a specified time limit for a real-time decision-making process.

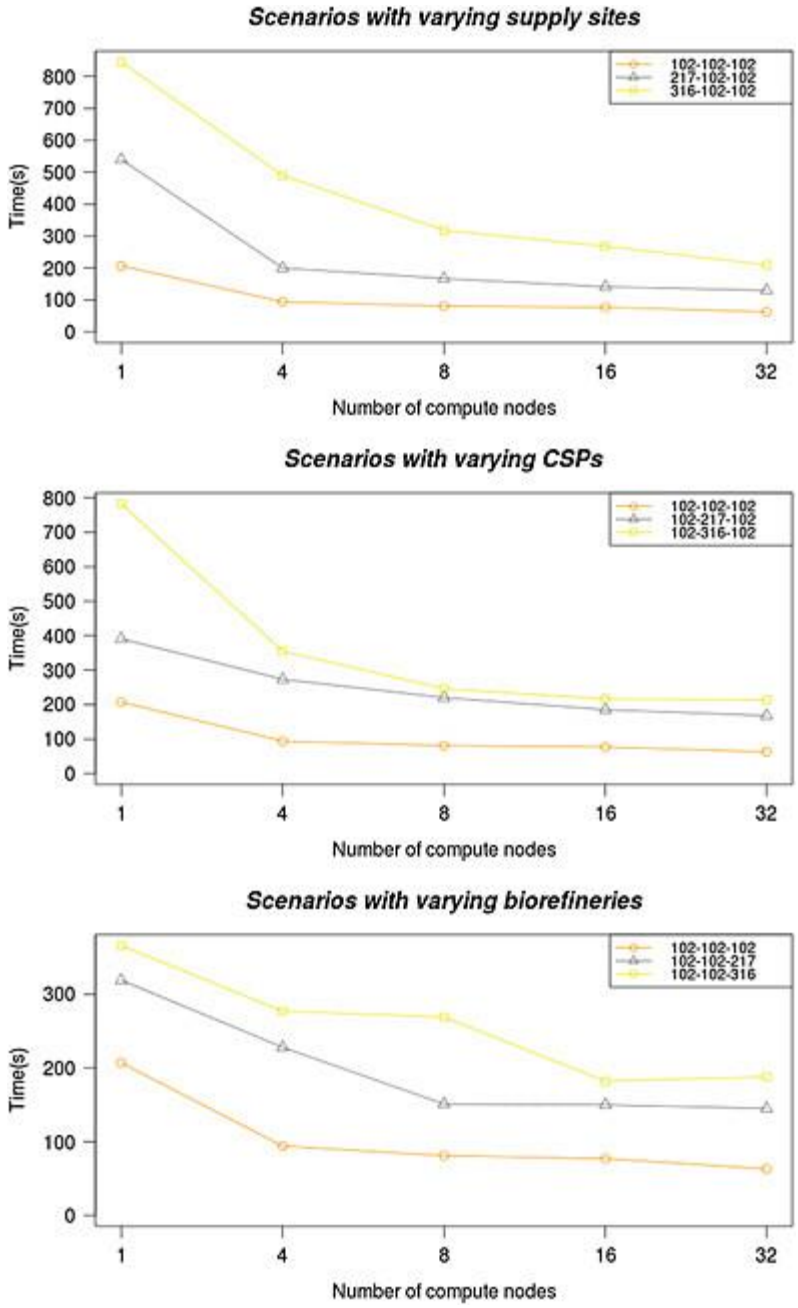


Figure 2.5. Performance profiles when fixed bioethanol demand (300 million gallons) and cropland usage rate (3%) are selected at various problem sizes. The three numbers in the legend represent number of candidate supply sites, candidate CSPs, and candidate biorefineries respectively.

The execution time of an MILP problem is often difficult to estimate theoretically because an MILP solver often combines a variety of combinatorial optimization strategies simultaneously in a single problem solving process and each strategy exhibits different computational intensity

(Williams, 2009). Although some general indicators, such as number of constraints and variables, are used to measure the difficulty of the MILP problem, the execution time of an MILP instance is heavily dependent on the problem structure. Therefore, the computational performance of the BioScope model is estimated using performance profiling. Our goal is to measure the weak scaling performance of the BioScope model and the underlying parallel MILP solver (i.e., Gurobi) in leveraging multiple compute nodes to solve more complex problem instances. Since randomization is used by the parallel MILP solver in building solution search trees and can greatly vary the total amount of numerical work to be done, speedup is not a straightforward performance measure (Liu et al., 2014). We thus measured the execution time of each experiment run to capture the relationship between the number of compute nodes used (each node uses all of the processor cores on it) and the amount of numerical work indicated by the selected model parameters.

Our performance profiling experiment considered two parameter sets: biomass (i.e., bioethanol demand and cropland usage rate) and spatial. In spatial parameters, we define the spatial problem size as a vector of parameters including candidate biomass supply sites, CSP facilities and biorefinery facilities. These parameters directly determine the number of constraints and variables in the formulated MILP problem. Accordingly, two types of performance profiling experiments are considered: 1) type I experiment: spatial scenarios with fixed bioethanol demand and cropland usage rate but changing problem size, and 2) type II experiment: biomass scenarios with fixed problem size but varying bioethanol demand and cropland usage rate. Type I experiment represents decision-making scenarios considering different locations to satisfy a fixed bioethanol production requirement. Type II experiment scenarios are often applied in sensitivity analysis of bioethanol demand and cropland usage rate considering a specified spatial context, e.g., a state. In general, model complexity increases with higher bioethanol demand or a larger spatial problem size, but

decreases with a higher cropland usage rate. For each experiment, up to 32 compute nodes (1,024 processor cores) on the Trestles supercomputer at San Diego Supercomputer Center were used. The execution time of all the runs ranged from 63 seconds to 28.7 minutes.

In type I experiment, the spatial parameters are determined by Table I. Starting from a base case scenario that only considers candidate supply sites and facility locations in Illinois (102 counties), six additional scenarios are included with extra candidates in Missouri (114 counties and one independent city) and Iowa (99 counties) for the selection of three spatial parameters: supply sites, CSPs and biorefineries. Tuning the spatial parameters in the seven scenarios directly change the MILP problem size as indicated in Table I. Results (Figure 2.5) show that when using a specified number of compute nodes, increasing the value of any of the three spatial parameters led to longer execution time of the model run. When more compute nodes were used, the model execution time was reduced attributing to the benefit of using parallel search strategies in the MILP solver. However, there is a limited improvement for model execution time after allocating more than 16 or 32 compute nodes. To explain this phenomenon, we broke down the time for all the model runs and found that the time spent on inter-node communication becomes a bottleneck when more compute nodes are allocated. As the number of compute nodes increased from 4 to 32, the percentage of time spent on communication and synchronization increased from 28.4% to 55.3%. As a result, the performance improvement becomes less significant when the compute nodes used is larger than 32.

Table 2.1 Problem size indicators for each scenario in Type I experiment

	Different Spatial Parameters			MILP Problem Size Indicators		
	Supply sites	CPSs	Biorefineries	# of constraints	# of variables	# of binary variables
Base case scenario	IL(102)	IL(102)	IL(102)	12,961	22,956	816
Varying supply sites	IL-MO(217)	IL(102)	IL(102)	24,806	34,686	816
	IL-MO-IA(316)	IL(102)	IL(102)	35,003	44,784	816
Varying CSPs	IL(102)	IL-MO(217)	IL(102)	26,186	47,681	1,276
	IL(102)	IL-MO-IA(316)	IL(102)	37,571	68,966	1,672
Varying biorefineries	IL(102)	IL(102)	IL-MO(217)	14,226	35,836	1,276
	IL(102)	IL(102)	IL-MO-IA(316)	15,315	46,924	1,672

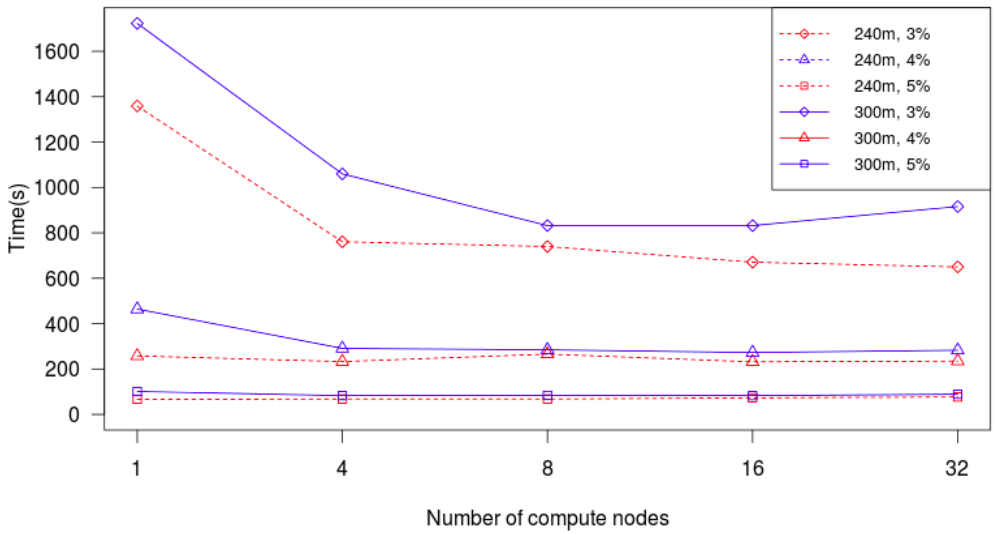


Figure 2.6. Performance profiles with six bioethanol demand and cropland usage rate settings at fixed problem size (316 candidate supply sites, 217 candidate CSPs and 102 candidate biorefineries).

In type II experiment, six typical combinations of bioethanol demand and cropland usage rate values are selected by taking two bioethanol production demand values at 240 and 300 million gallons, and three cropland usage rate at 3%, 4%, and 5%. Results (Figure 2.6) show that scenarios

with less cropland usage rate (e.g., 3%) took longer to compute and increasing the number of nodes can significantly reduce the execution time. Under the same cropland usage rate, larger bioethanol demands tend to introduce larger computing cost. The execution time of those scenarios decreased as more nodes were added. On the other hand, scenarios with higher cropland usage rate (greater than 3%) and lower bioethanol demand (less than 300 million gallons) do not show significant computational performance improvement when more compute nodes are added (Figure 2.6). This can be explained by the underlying parallelism implemented by Gurobi. The distributed MILP in Gurobi consists of two phases: concurrent phase and distributed phase. In the concurrent phase, different strategies (e.g., randomized search) are sent to each compute node to build search trees until the problem is solved or enough search nodes are generated. A winner that has made the most progress is selected for the distributed phase in which the search continues by dividing the partially explored MILP search tree of the winner to available compute nodes to solve in parallel. The concurrent phase involves multiple strategies and the randomization of the run of each strategy, which is less scalable to the number of compute nodes. The distributed phase is more scalable by simply dispatching search nodes to compute nodes. Higher cropland usage rate and lower bioethanol demand often lead to less complex MILP instances. For these instances, the concurrent phase can produce near- optimal or even the optimal solution. The distributed phase, thus, did not contribute much to the model execution.

In summary, the two experiments drew useful guidelines for allocating computing resources to the biomass and spatial parameter sets, respectively. Type I experiment clearly showed the relationship between the computational intensity and spatial parameters of the BioScope model. Type II experiment showed that the combination of bioethanol demand and cropland usage rate are two important indicators for resource allocation when considering biomass parameters in the

model. Different value ranges for these two parameters may lead to different resource allocation strategies: instances with less cropland usage rate may benefit more from parallel computing; while higher cropland usage rate and lower bioethanol demand may not need many compute nodes. More work is undergoing to gain additional insights of the parallelization strategies of the MILP solver and how to reduce the communication overhead when a large number of compute nodes are used. We are integrating our findings obtained from type I results into GISolve and CyberGIS Gateway for guided resource allocation and parameter selection.

2.5 CONCLUSION AND FUTURE WORK

BioScope, a supply chain optimization model, is integrated in the CyberGIS Gateway as a science gateway application. The application follows a geodesign approach for spatial decision-making, environment design, and a service integration approach provided by the GISolve middleware services. It provides a highly interactive user environment for community users and supports the optimization of biomass supply chain and the evaluation of uncertainty through a lightweight computational workflow implementation. The development of computation workflow for uncertainty analysis leverages the job dependency feature of existing job schedulers on cyberinfrastructure resources and s ignificantly simplifies the execution and management of cyberGIS analytics. The proposed CyberGIS-BioScope can be served as a guideline for future development of discipline-specific problem-solving environments that requires spatial data access and sharing, intensive computation and modeling, and interactive decision-making.

The target user groups of the CyberGIS-BioScope application include decision makers, like design engineers, investors, and policy makers, and educators. The CyberGIS-BioScope application provides a comprehensive support at all stages from model development, scenarios

analysis,and evaluation to the final decision-making. Benefiting from the transparent access to data, models,and advanced cyberinfrastructure, community users could speed up their scientific discovery and knowledge-sharing process in the collaborative biomass-to-biofuel supply chain research.

BioScope is being extended to incorporate multiple cross-domain models to enable further understanding of the factors that influence the supply chain systems, including weather models (e.g., Weather Research Forecasting model, http://wrf-model.org) and multiple crop simulation models (e.g., BioCro, Miguez et al., 2012).Thus, CyberGIS-BioScope will be extended to accommodate more sophisticated model computation and spatial decision-making components. The data access, integration and processing, and the complexity of job dependencies will be introduced by linking multiple models from cross-domain.We are exploring existing workflow tools such as Apache Airavata (http://airavata.apache.org) and Kepler (http://kepler-project.org) to better scale the workflow execution to multiple cyberinfrastructure resources.

CHAPTER 3
A CYBERGIS APPROACH TO UNCERTAINTY AND SENSITIVITY ANALYSIS IN BIOMASS SUPPLY CHAIN OPTIMIZATION

This chapter extends CyberGIS-BioScope, a cyberGIS-enabled decision support system developed in Chapter 2, to enable the optimization of biomass feedstock provision under uncertainties. Particularly, a cyberGIS approach is proposed with the integration of data management, mathematical modeling, uncertainty and sensitivity analysis, what-if scenario analysis, and result representation and visualization. Leveraging high-performance computing capabilities provisioned by advanced cyberinfrastructure, the proposed cyberGIS approach enables Monte Carlo based uncertainty and sensitivity analysis for optimization modeling by resolving significant computational intensity. This approach generalizes and streamlines data management, computation, and visualization components, so it is expected to work on biomass supply chain optimization applications customized by different model and data inputs.

This chapter cannot be realized without successful teamwork. The team members include Hao Hu, Tao Lin, Shaowen Wang and Luis Rodríguez. Mr. Hu and Dr. Lin together led the overall research design, result discussion, and manuscript writing. The primary tasks conducted by Mr. Hu include 1) studied and synthesized existing methods in uncertainty and sensitivity analysis; 2) integrated the module of interactive scenario analysis, uncertainty and sensitivity analysis into the CyberGIS-BioScope system; and 3) led the data collecting, processing, visual analytics and drove the development of conclusions. Dr. Lin contributed significantly to the manuscript writing and

revision. Drs.Rodríguez and Wang participated in the overall research design, results discussions, and the draft revision. The content of this chapter has been published in a journal paper (Hu et al., 2017).

Abstract. *Decision making in biomass supply chain management is subject to uncertainties in a number of factors such as biomass yield, procurement prices, market demands, transportation costs, and processing technologies. To better understand such uncertainties requires statistical analysis and data-intensive computing enabled by cyberGIS (aka geographic information science and systems based on advanced cyberinfrastructure and e-science). Therefore, we have developed a cyberGIS approach to optimize biomass supply chains under uncertainties. Our approach 1) designs optimal biomass supply chains from regional to national scale with flexible spatial selection of study areas; 2) performs uncertainty and sensitivity analysis to quantify how various sources of uncertainty in the biomass supply chain contribute to the variation of optimal results; and 3) provides users with online geodesign features. This approach has been implemented as a decision support system through integration of data management, mathematical modeling, uncertainty and sensitivity analysis, scenario analysis, and result representation and visualization. An optimization modeling analysis of 7,000 scenarios using Monte Carlo methods has been conducted to quantify the uncertainty and sensitivity impact of various input factors on ethanol production costs and optimal biomass supply chain configurations in Illinois, United States. The results from uncertainty analysis showed that the minimal ethanol production costs range from \$2.30 to \$3.43 gal^{-1}, considering uncertainties from biomass supply, transportation, and processing. The results of sensitivity analysis demonstrated that biomass-ethanol conversion rate was the most influential factor to ethanol production costs while the optimal biomass supply chain*

infrastructure was sensitive to changes in biomass yield, raw biomass transportation cost, and logistics loss rate. Leveraging high performance computing power through cutting-edge cyberGIS software, what-if scenario analysis has been evaluated to make decisions in case of unexpected events occurring in the supply chain operations.

Keywords. Biomass supply chain; CyberGIS; Spatial decision support system; Geodesign; Optimization; Uncertainty and sensitivity analysis

3.1 INTRODUCTION

Spatial decision making often applies geographic information science and systems (GIS) to integrate spatial data and various models to evaluate policy and decision options at different geospatial scales. It has become important to many domains including agriculture, climate, land use, and transportation. As the value and sustainability of renewable energy have been widely recognized, spatial decisions associated with the development of renewable fuels has received increasing attention in recent years. Biofuels can help reduce greenhouse gas emissions, improve energy security, and spur rural economies (Tilman et al., 2009). In additional to the well-established biofuel production from corn and sugarcane, recent studies focus on cellulosic biofuel development, produced from non-food biomass sources such as agricultural and woody residuals, to enhance food and energy security (Balat and Balat, 2009; Srirangan et al., 2012). How to achieve efficient and effective designs of the biomass-to-bioenergy supply chain is critical for large-scale cellulosic biofuel production (Hess et al., 2007; Lin et al, 2013; Sharma et al., 2013).

Biomass supply chains consist of multiple stages, from biomass production, harvesting, preprocessing, conversion, and distribution to the eventual end use of biofuels. This study focuses on upstream biomass supply chains, which typically includes biomass suppliers, storage and pre-

processing sites, biorefinery, and transportation (Lin et al, 2013; Vlachos et al., 2008; Becher and Kaltschmitt 2013; Yadav 2016). Key decisions within biomass supply chains include biomass provision plans, storage and transportation strategies, and the number, location, and capacity of facilities for biomass processing and conversion. A number of mathematical programming models have been developed to optimize biomass supply chain configurations with case studies in different countries including the United States (Lin et al, 2013; Marvin et al., 2012; You et al., 2012; Zhang et al., 2013; Lin et al., 2014), Spain (Panichelli and Gnansounou, 2008), Germany (Nagel, 2000), Greece (Rentizelas et al., 2009), Finland (Höhn et al., 2014), and Brazil (Jonker et al., 2016), while most optimization models are deterministic, without considering uncertainties in system inputs.

Uncertainty, however, is inevitable for the development of biomass supply chains, given the nature of bioenergy production and technological improvement. Changes in biomass availability and yield, market demand, transportation costs, and processing technology would significantly affect overall system configurations. Decision makers are interested in learning how uncertainties in these factors contribute to the design of biomass supply chains, and where to make a change among those factors to improve optimal biomass supply chain decisions, especially with the capability of data-driven analytics. Previous studies (Lin et al, 2013; Marvin et al., 2012; Parker et al., 2010) have employed sensitivity analysis to understand the impact of specific input variables on the optimal biomass supply chain. However, this sensitivity analysis approach is limited in situations where various sources of uncertainty need to be addressed together, and their propagation to decisions have to be understood systematically.

Uncertainty and sensitivity analysis techniques are often used to understand how decision outcomes respond to changes in inputs (Lilburne and Tarantola, 2009). Uncertainty analysis aims at understanding the variability of outcomes given model input uncertainty, while sensitivity

analysis focuses on quantifying the impact of each input that is responsible for this variability. Local sensitivity analysis serves to quantify the rate of output change relative to the change of each individual input parameter. Considering potential effects from simultaneous variations of multiple inputs, global sensitivity analysis identifies the most important input parameters and quantifies the contributions of various parameter subsets to the variation of the overall decision outcome. Methods for uncertainty and sensitivity analysis require a large number of model evaluations. As a result, enough information is collected to support distribution function analysis of model outcomes. In the meantime, computational efficiency and feasibility pose a major bottleneck, especially for optimization modeling approaches that are already computationally intensive.

Oftentimes, the optimization of biomass supply chains should be considered as dynamic management decisions in response to uncertain situations occurring in the future. For example, an established biomass supply chain is subject to uncertainties such as shortages in biomass supply, changes in transportation costs, economic competition, and increases in bioethanol demand. Decision makers need to understand how these changes affect the operation of biomass supply. Therefore, interactive and exploratory decision support is necessary and desirable.

Due to the spatial nature of biomass distribution, logistics, and facility location decisions, previous work has demonstrated integration of GIS with domain-specific models for studying biomass potential (Voivontas et al., 2001; Fiorese and Guariso, 2010), logistics and production costs (Panichelli and Gnansounou, 2008; Höhn et al., 2014; Graham et al., 2000), and supply chain management (Lin et al, 2013; Zhang et al., 2013; Frombo et al., 2009]. By taking advantage of spatial data management, processing, analysis, and visualization capabilities provided by GIS, biomass supply chain optimization support could become accessible by decision makers. Sensitivity and what-if analysis are generally required in complex decision problems. With the

support of enabling technologies in Geodesign (Steinitz, 2012), GIS-based decision support systems allow users to engage in what-if scenarios to quickly and easily determine the impacts of different problem settings. However, conventional GIS approaches to decision support for biomass supply chains are limited to small-scale problems due to lack of computational scalability and collaboration support for decision making.

CyberGIS is based on advanced cyberinfrastructure for achieving computational scalability and collaborative problem solving, and has emerged as a new-generation GIS in the era of big data (Wang, 2010; Wang et al., 2013; Wang et al., 2015; Wang, 1017). Leveraging high-performance computing power provisioned by advanced cyberinfrastructure, cyberGIS can provide user-friendly online capabilities for spatial decision support (Hu et al., 2015). Through seamless integration of cyberGIS and optimization modeling, a CyberGIS-BioScope application has been developed to enable online decision support for deterministic biomass supply chain optimization based on scalable computing capabilities (Lin et al., 2015). As decision makers are increasingly interested to understand uncertainties, capabilities for uncertainty quantification and sensitivity measurement need to be integrated into decision support tools.

The objectives of this study are to: 1) quantify the impact of uncertainty and sensitivity on biomass supply chains using Monte Carlo methods; 2) develop a novel cyberGIS approach to enable computing-intensive uncertainty and sensitivity analysis; and 3) establish interactive and user-friendly geodesign capabilities for decision makers to conduct biomass supply chain optimization under uncertainties. The proposed cyberGIS approach is implemented by providing decision support for county-level biomass supply chain optimization problems with a specific consideration of Miscanthus as the cellulosic biomass feedstock. The results of uncertainty and sensitivity analysis are discussed based on related case studies in the State of Illinois.

3.2 METHODS

3.2.1 Uncertain Factors in Biomass Supply Chain Optimization

A typical biomass provision system may include three stages: farms, centralized storage and preprocessing sites (CSPs), and biorefineries. The BioScope model was developed to optimize strategic planning decisions to support effective biomass feedstock provision for large-scale ethanol production (Lin et al., 2013). The model is a mixed integer linear programming (MILP) model that minimizes annual ethanol production costs. Annual ethanol production costs (Z) consist of four components: biomass purchase costs (CB), transportation related costs (CT), CSP site related costs (CS), and biorefinery related costs (CE) (Eq. (3.1)),

$$Minimize\, Z = C_B + C_T + C_S + C_E \tag{3.1}$$

The BioScope model requires inputs from biomass yield, cropland usage rate, transportation distances, transportation costs, that consist of distance variable transportation cost and fixed transportation cost for loading and unloading, logistics related losses due to handling, biomass-ethanol conversion rate, and facility related costs. The detailed description of the objective function and constraints of the BioScope model are not provided in this study, but can be found in (Lin et al., 2013). By considering constraints from biomass supply, transportation, to biofuel production, the BioScope model aims to minimize annual ethanol production costs by optimizing the following decision variables:

- The number, capacity, and location of supply sites
- The number, capacity, and location of CSPs
- The number, capacity, and location of biorefineries
- Amount of raw biomass transported from supply sites to CSPs for preprocessing

- Amount of preprocessed biomass transported from CSPs to biorefinery for biofuel conversion

The key uncertain factors associated within a typical three-stage biomass supply chain include biomass yield and supply, transportation costs, logistics related losses due to handling, and improvement of conversion efficiency (Figure 3.1). Given the variations of weather and farmers' participation, it is expected that biomass yield and cropland usage rate will vary. Transportation costs are largely dependent on the gasoline price that has changed significantly in the recent years. Biomass-ethanol conversion technologies are under active development, where the conversion rate varies with technology and should be expected to vary again with scale-up production. The variations of these uncertain factors would cause changes in the optimal design of the biomass provision system and associated production costs.

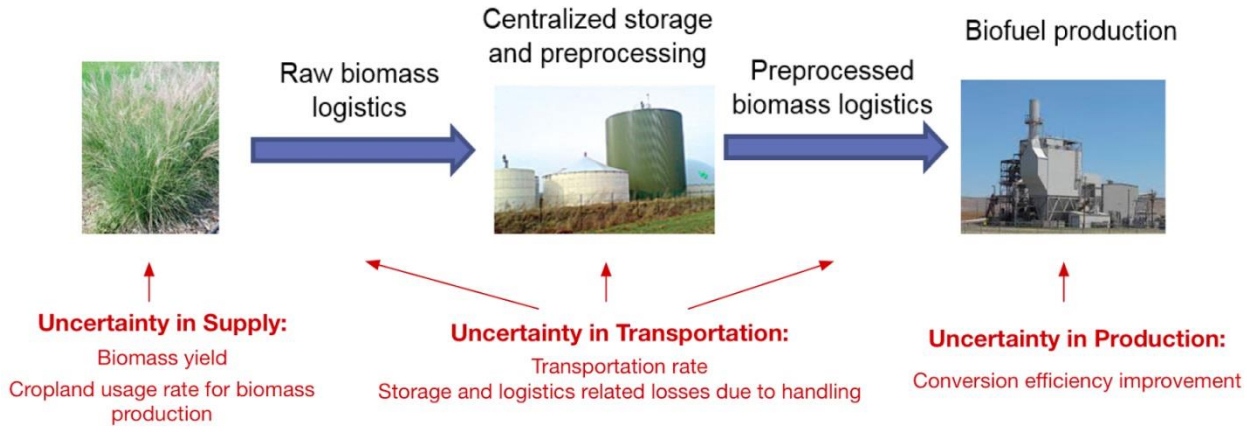


Figure 3.1. Uncertainties involved in a three-stage biomass supply chain analysis.

3.2.2 Monte Carlo based Uncertainty and Sensitivity Analysis

Monte Carlo based methodologies have been widely used to perform uncertainty and sensitivity analysis when a model is too complex to be estimated analytically (Crosetto et al., 2000). Multiple model evaluations are simulated with randomly selected model inputs, and uncertainty and

sensitivity analysis can be further performed based on these evaluations. By sampling from a full range of possible inputs, contribution to the variation of outputs are evaluated using global sensitivity indices, including for example Sobol's indices and related variants (Sobol, 1990; Sobol, 2001; Saltelli, 2002; Saltelli et al., 2010), Fourier Amplitude Sensitivity Test (FAST) indices (Cukier et al., 1978), and extended FAST (Saltelli et al., 1999). A complete Monte Carlo based methodology, including sensitivity analysis consists of the following steps:

1) Select a model and identify key input parameters for analysis;

2) Use a sampling approach to generate scenarios from the selected inputs based on the given range and distribution;

3) Run the model based on the input scenarios generated in Step 2;

4) Perform uncertainty and sensitivity analysis using all the results of model output in Step 3.

How to sample effectively from high dimensional parameter spaces to represent the entire study space is critical when considering uncertainties from multiple parameters simultaneously. A quasi-random sequence was applied in this study because it was designed to generate samples as uniformly as possible over high dimensional spaces and it has been proved to outperform random numbers when using Monte Carlo based approaches for sampling the estimation of high dimensional integrals (Saltelli et al., 2010; Sobol and Kucherenko, 2005).

The distribution of model output can be summarized statistically for uncertainty analysis. Variance-based decomposition was applied to conduct sensitivity analysis by quantifying the proportion of the output variance contributed by each input. Considering the model,

$$Y = f(X_1, X_2, X_3, ..., X_k) \tag{3.2}$$

where X_i are independent inputs, and Y are scalar response variables. Variance-based decomposition can estimate how uncertainty in Y can be apportioned to different sources of inputs X_i,

$$V(Y) = \sum_i V_i + \sum_i \sum_{i<j} V_{ij} + \sum_i \sum_{i<j} \sum_{j<l} V_{ijl} + \ldots + V_{1\ldots k} \tag{3.3}$$

where V_i is the proportion of output variance contributed by the *i*th model input (not be confused with the variance of input X_i) and it represents the sensitivity of Y to X_i, V_{ij} is the second order term showing the proportion of the output variance contributed by the interaction between the *i*th and *j*th inputs (similarly, not the covariance of input X_i and X_j) , V_{ijl} and $V_{1\ldots k}$ have similar interpretations but with higher order of interactions, *k* is the total number of model inputs. The first-order (Eq. (3.4)) and total-effect sensitivity index (Eq. (3.5)) are often estimated in variance-based sensitivity analysis.

$$S_i = \frac{V_i}{V(Y)} \tag{4}$$

$$S_{Ti} = 1 - \frac{V_{-i}}{V(Y)} \tag{5}$$

where V_{-i} is the proportion of the output variance contributed by all other model inputs and their interactions except X_i. The First-order index (S_i) measures the effect of varying X_i alone, while the total-effect index (S_{Ti}) includes the effect of varying X_i alone and with other input variables of any high order interactions.

There are a number of methods to estimate S_i and S_{Ti} in literature (Sobol, 1990; Sobol, 2001; Saltelli et al., 2010; Jansen, 1999; Homma and Saltelli, 1996). Among all the variants, the first-order sensitivity index in Saltelli et al. (2010) was selected in this study given that there is no clear conclusion drawn from literature that any first-order sensitivity index outperformed the others.

Jansen's total effect (Jansen, 1999), on the other hand, requires the fewest model runs to be estimated but achieves competitive approximations compared with other estimators (Saltelli et al., 2010), and therefore was selected in this study. Saltelli's first-order and Jansen's total-effect sensitivity index are denoted as S_i and S_{Ti} in this study (Eq. (3.6) and Eq. (3.7)).

$$\text{First-order sensitivity index: } S_i = \frac{1}{N}\sum_{j=1}^{N} f(\boldsymbol{B})_j \times (f(\boldsymbol{A}_{\boldsymbol{B}}^{(i)})_j - f(\boldsymbol{A})_j)/V(Y) \tag{3.6}$$

$$\text{Total-effect sensitivity index: } S_{Ti} = 1 - \frac{1}{2N}\sum_{j=1}^{N} (f(\boldsymbol{B})_j - f(\boldsymbol{A}_{\boldsymbol{B}}^{(i)})_j)^2 / V(Y) \tag{3.7}$$

where $\boldsymbol{A}$ and $\boldsymbol{B}$ are matrices formed by two independent samples of model inputs, with each row indicating the sample index and each column representing the input variables. $\boldsymbol{A}_{\boldsymbol{B}}^{(i)}$ represents the derived matrix where column *i* is selected from matrix $\boldsymbol{B}$ and replaces the corresponding column in $\boldsymbol{A}$. $(\boldsymbol{B})_j$ denotes *j*-th row of matrix **B** and $\mathbf{f}(\boldsymbol{B})_j$ can be evaluated to determine *Y* at row *j* as in Eq. (2). In this study, parameters in matrix $\boldsymbol{A}$ and $\boldsymbol{B}$ are generated based on Monte Carlo sampling methods with a quasi-random sequence. More details about matrices construction are provided in Appendix A. According to Eq. (3.6) and (3.7), the model needs to be evaluated using all parameters in $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{A}_{\boldsymbol{B}}^{(i)}$, but $\boldsymbol{B}_{\boldsymbol{A}}^{(i)}$. Given *k* input parameters under analysis, $\boldsymbol{A}_{\boldsymbol{B}}^{(i)}$ need to be evaluated $N \times k$ times. As a result, the total number of model evaluation is $N \times (k + 2)$, which includes 2*N* evaluations for $\boldsymbol{A}$ and $\boldsymbol{B}$. As a common practice, the convergence of results requires the independent sampling size *N* to be sufficiently large, for example 500 or higher as suggested by (Saltelli et al., 2010).

The set of $N \times (k + 2)$ scenarios only generates one first-order sensitivity index and one total-effect sensitivity index. Therefore, the estimated sensitivity indices may vary from different sampling sets since the sampling procedure itself is stochastic. As it would be too costly to repeat the simulation of multiple sets of $N \times (k + 2)$ scenarios, a better strategy adopted in this study is

to assess confidence intervals using bootstrapping with a resampling method (Efron and Tibshirani, 1993). Specifically, N samples used for the model simulation were resampled 100 times with replacement, and then form 100 extra sets of $N \times (k+2)$ scenarios. Since scenarios in 100 extra sets are all resampled, it is not necessary to conduct extra scenario simulation. In this way, distributions for S_i and S_{Ti}, are obtained and the 95% confidence intervals are constructed with no extra computational cost.

Cellulosic-based biomass supply chain systems undergo changes given the variations of natural and economic factors and the development of novel conversion technologies. To represent the uncertainties in a typical three-stage biomass supply chain (Figure 1), five key parameters were selected for this study. These includes biomass yield rate, cropland usage rate, raw biomass transportation cost, logistics loss rate, and biomass-ethanol conversion rate. Raw biomass transportation cost is referred to as the distance variable cost for transporting raw biomass. It is affected by bale density of raw biomass feedstocks. The description for each parameter is provided in Table 3.1.

Based on historical data and the literature review, we identified the variation ranges of each parameter to better cover the uncertainty space for the biomass supply chain study (Table 3.1). Since no prior probability distribution information is available, a uniform distribution was assumed within the parameter space. By drawing 1,000 independent samples from the assumed parameter spaces of the five inputs, we ran the Monte Carlo process based on quasi-random sequence to generate 7,000 (i.e. 1,000×(5+2)) model scenarios for quantification of how various sources of uncertainty contribute to the optimal design of biomass supply chain.

Table 3.1 – The ranges of possible values of input parameters used in uncertainty analysis.

Variable	Description	Lower	Baseline	Upper
Yield rate*	A coefficient term applied to adjust biomass yields predicted by MISCANMOD model in (Jain et al., 2010)	0.75 (-25%)	1	1.25 (+25%)
Cropland usage rate	The percentage of land that can be potentially converted to grow bioenergy crop	3% (-40%)	5%	7% (+40%)
Raw biomass transportation cost **	Variable costs that increase linearly with distance for transporting raw biomass (in $ $Mg^{-1} km^{-1}$)	0.12 (-20%)	0.15	0.18 (+20%)
Logistics loss rate	A coefficient term applied to storage and logistics loss related to handling	3% (-40%)	5%	7% (+40%)
Conversion rate (gal/Mg)***	The efficiency of biomass to bioethanol conversion	66 (-15%)	78	90 (+15%)

* The base case of county-level Miscanthus yield rate in Illinois is based on MISCANMOD model (Jain et al., 2010)

** The range of raw biomass transportation cost is based on (Lin et al., 2013)

***The range for biomass-ethanol conversion rate is based on (Humbird et al., 2011) and (Wyman, 2007)

3.2.3. CyberGIS-BioScope Decision Support System

CyberGIS-BioScope decision support system was developed to provide user-friendly BioScope and geospatial analytical capabilities while achieving scalable optimization problem solving and collaborative decision support (Lin et al., 2015). The architecture of the CyberGIS-BioScope system includes three major components: cyberinfrastructure, cyberGIS tools and services, and a web-based user interface (Hu et al., 2015). CyberGIS-BioScope exploits high-performance cyberinfrastructure resources to enable complex optimization model solving and computationally intensive simulations for uncertainty and sensitivity analysis through modular and open-service architecture (Figure 3.2). This study has a particular focus on creating new services for uncertainty and sensitivity analysis. For example, a quasi-random sampling service was established to generate Monte Carlo based scenario configurations. Services for visual analytics were developed to

support the evaluation of uncertainty analysis results. Our system provides user-friendly interfaces that are capable of group decision support, scenario design, spatial query, and map-based visualization (Figure 3.3). A typical workflow for uncertainty and sensitivity analysis in CyberGIS-BioScope consists of the following steps:

1. In the *Scenario Design* component, users specify the range and distribution of the uncertain model parameters from the user interface;
2. Based on the information from Step 1, the quasi-random sampling service in the *Uncertainty and Sensitivity Analysis Tool* component is called to generate large number of Monte Carlo based scenario configurations;
3. The *Biomass Supply Chain Model and Solver* component takes the scenario configurations and constructs batch jobs that are ready for submitting to the backend cyberinfrastructure;
4. The *Computation Management* triggers the batch job submission and ensures as many as jobs are running in parallel given the high throughput capability of the backend;
5. Model results are sent back to the *Uncertainty and Sensitivity Analysis Tool* component to perform uncertainty and sensitivity measurements once all the jobs are completed;
6. The results of uncertainty and sensitivity analysis are presented as tables, charts as well as map visualizations on the user interface.

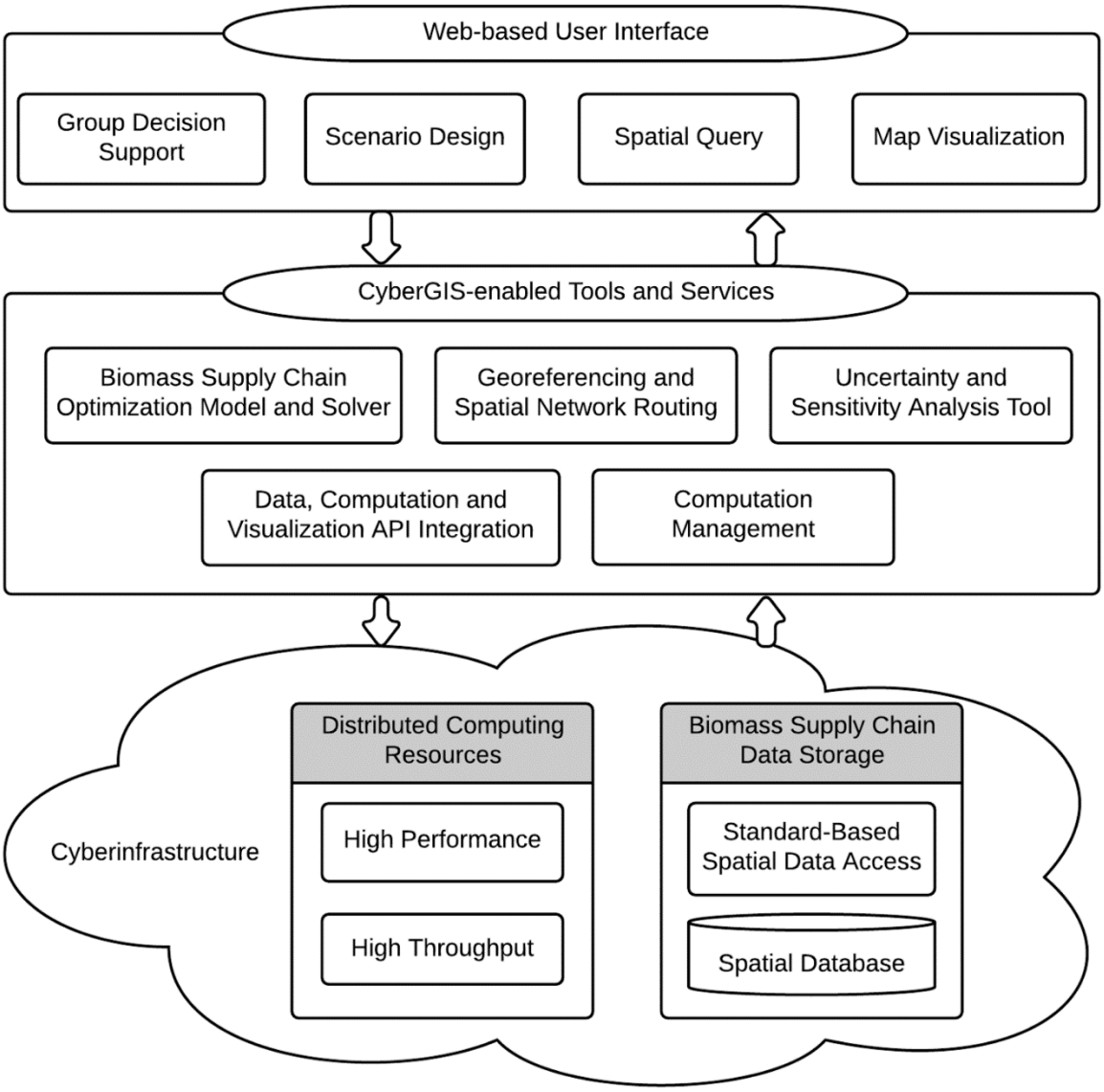


Figure 3.2 Architecture of CyberGIS-BioScope decision support system.

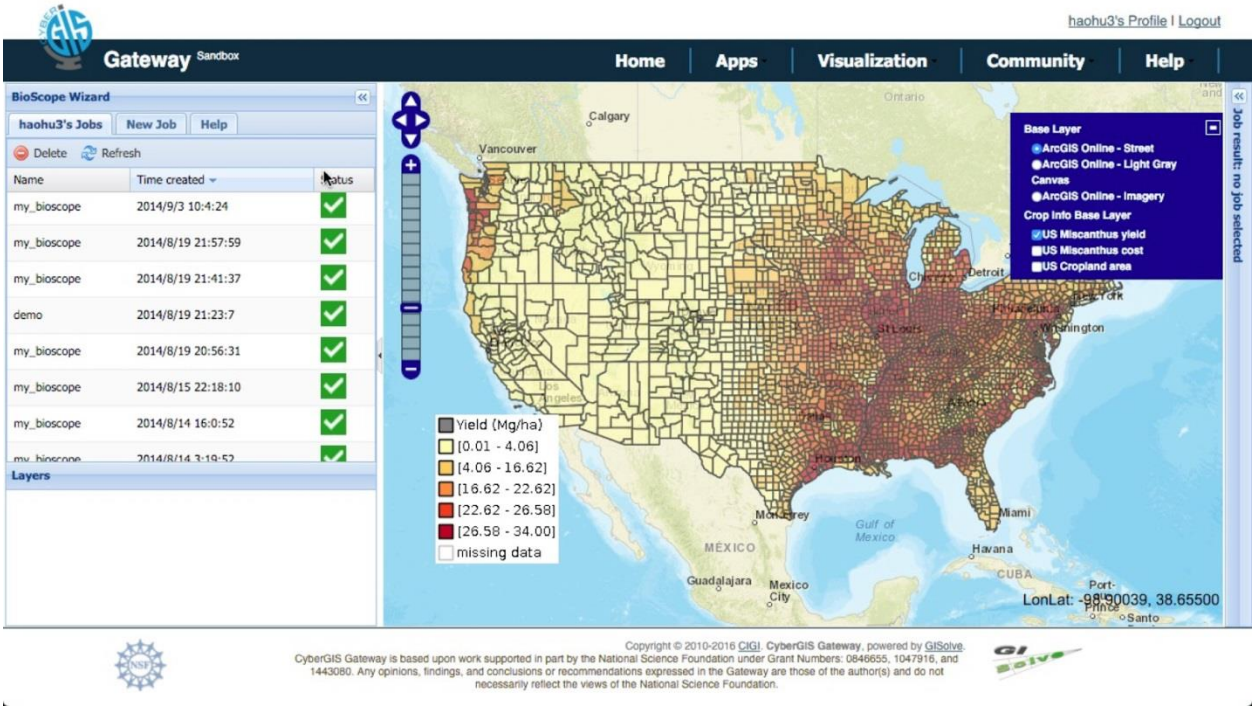


Figure 3.3 CyberGIS-BioScope user interface.

3.3 DATA AND CASE STUDY

To illustrate the application of the CyberGIS-BioScope decision support system, we performed a case study in the State of Illinois considering Miscanthus as the cellulosic biomass feedstock for county-level supply chain decision making. The CyberGIS-BioScope system manages county-level Miscanthus yield (Jain et al., 2010), cost, and cropland area (USDA, the Census of Agriculture) as base case information. Each county was assumed to have at most one CSP site and one biorefinery facility, which were expected to be located at the county seat. Biomass was assumed to move by road transportation. The shortest distance between each pair of counties was calculated using the Open Source Routing Machine (OSRM) (Luxen and Vetter, 2011). For all scenario analyses, annual ethanol production demand is fixed at 160,000,000 gal, which is approximately the capacity of a large biorefinery.

The case study consists of two parts. For the first part (Section 3.4), we focused on uncertainty and sensitivity analysis of the biomass supply chain given various sources of uncertain parameters. By considering the whole variation range of uncertain parameters, we would like to see all the possibilities of optimal biomass supply chain configurations as well as the level of influence caused by these of uncertain parameters. In the second part (Section 3.5), we demonstrated three what-if scenario analyses to show interactive and exploratory decision support capabilities enabled by cyberGIS, in case of unexpected events occurring in the supply chain operations. The results of case study were further discussed in Section 3.6 with an emphasis on how the proposed cyberGIS approach could be used for biomass supply chain decision making in real applications.

3.4 UNCERTAINTY AND SENSITIVITY ANALYSIS

3.4.1 Base Case Analysis

A base case was implemented with baseline values of Miscanthus yield rate, cropland usage, raw biomass transportation cost, logistic loss rate and biofuel conversion rate listed in Table 1. The results of base case show that the optimal ethanol production costs is \$2.713 gal^{-1}. Biorefinery related costs account for 49% of total costs, followed by biomass procurement cost (31%), CSP related cost (13%), and transportation cost (7%).

3.4.2 Uncertainty Analysis

According to the results of uncertainty analysis, the optimal ethanol production costs range from \$2.30 to \$3.43 gal^{-1} (Figure 3.4), with mean value \$2.77 gal^{-1} and standard deviation \$0.25 gal^{-1} (Table 3.2). In addition to the variation on ethanol production costs, the optimized biomass supply chain configurations also vary with different scenarios. The number of supply counties varies from

7 to 38 counties, whereas the number of CSP facilities varies from 1 to 11. All the scenarios choose to build one centralized biorefinery to achieve maximum economies of scale. The changes of biomass yield and cropland usage rate would affect the amount of biomass supply from each county. Given a fixed biomass demand, the variance of county-level biomass availability changes the required biomass sourcing area, which would further affect biomass transportation distances and costs. The increased raw biomass transportation cost and logistics losses will increase the minimal ethanol production costs, as a result of increased overall transportation costs and biomass amount to be processed.

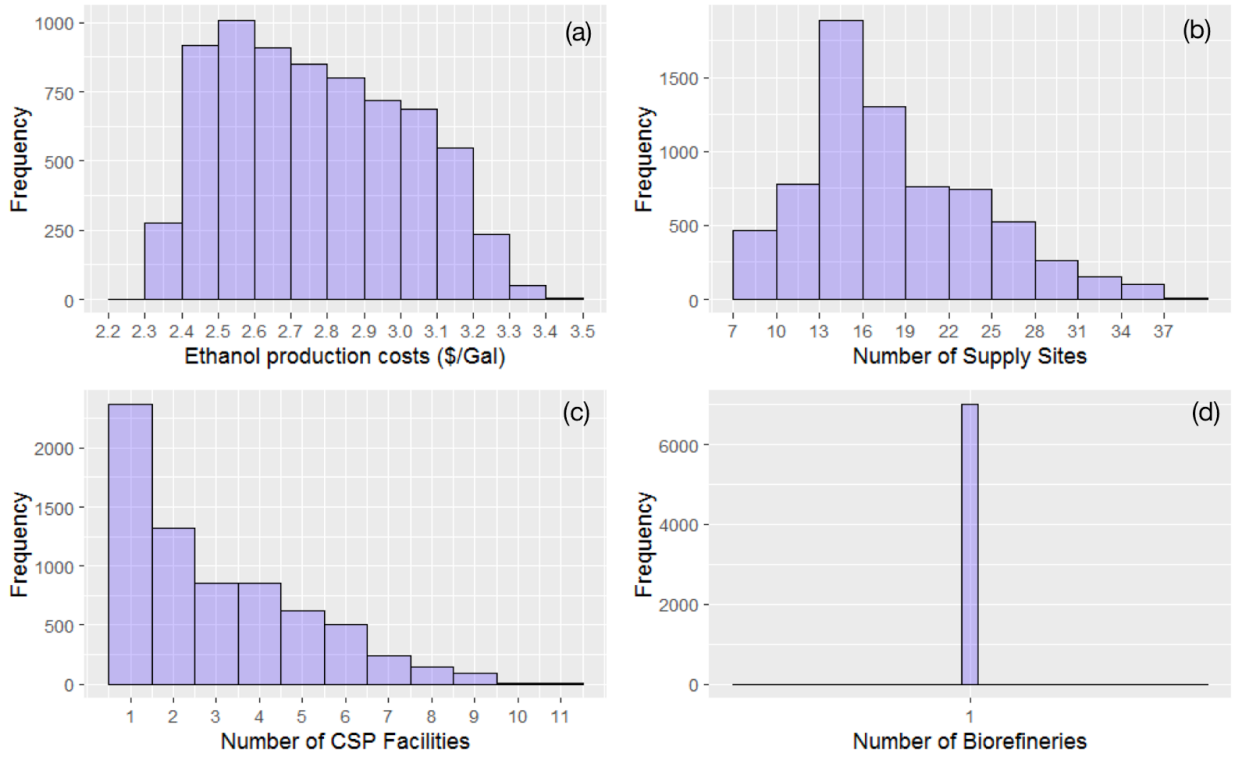


Figure 3.4. Histogram of key optimized biomass supply chain decision variables of 7,000 scenarios: (a) optimized ethanol production cost, ranging from $2.30 to $3.43 gal^{-1}; (b) number of supply counties, most scenarios select 13-16 counties; (c) number of CSP sites, nearly 30% scenarios select one CSP facility; and (d) number of biorefinery sites, all scenarios select one site.

The optimal supply chain configurations for the base case includes one biorefinery located in Jefferson County and three CSPs located in Washington, Jefferson, and Hamilton County, respectively (Figure 3.5: Left). Among 7,000 possible scenarios, southern Illinois counties should be selected to build a biomass supply chain system (Figure 3.5). Jefferson County has the highest possibility, above 60%, to be selected for CSP and biorefinery facilities. The variations of optimal CSP facilities are higher than biorefinery facilities, suggested by a wider range of locations potentially selected (Figure 3.5: Right). Twenty supply counties were selected to be biomass suppliers among these scenarios. The most likely biomass supply counties are always located surrounding to these highly possible CSP counties, attributing to their competitive advantages on transportation distances.

Table 3.2 The statistical summary of 7,000 scenario analyses

	Ethanol Production Costs ($ gal^{-1})	Number of Suppliers	Number of CSPs	Number of Biorefineries
Mean	2.7674	17.7	2.9	1
Standard Deviation	0.2488	6.1	2.0	0

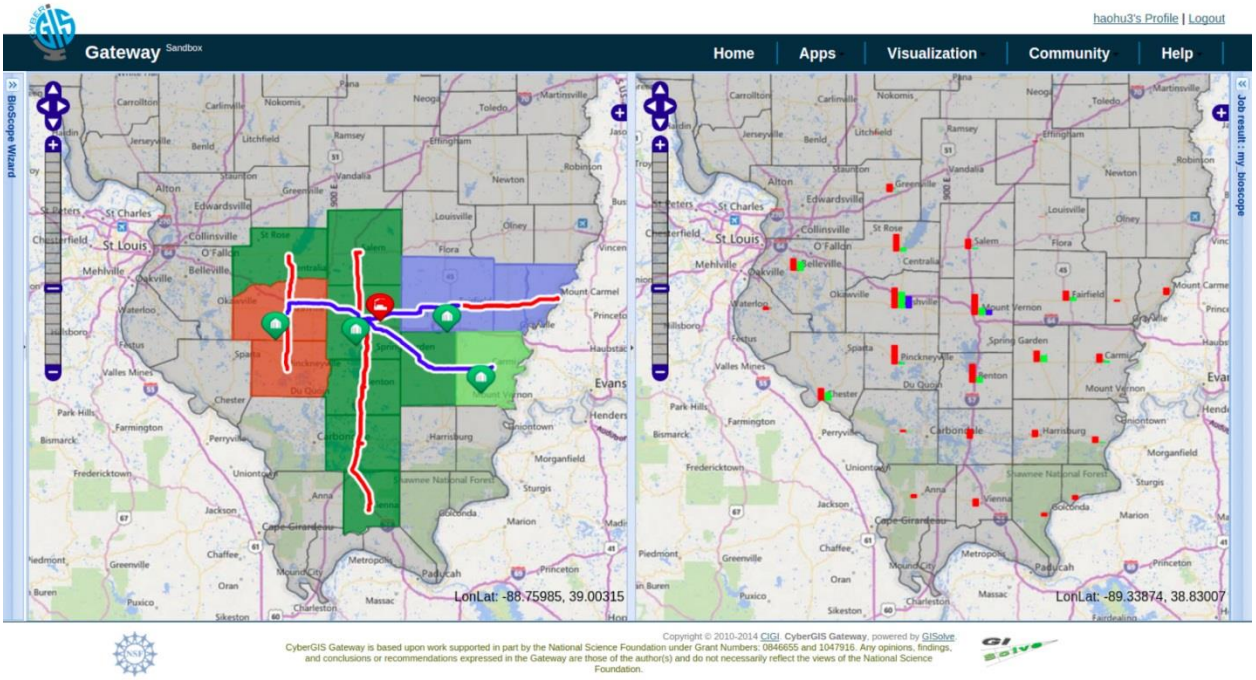


Figure 3.5. Biomass supply chain decisions in Illinois. Left: base case. Right: summary of 7,000 scenarios. On the right map, the bar charts represent the probability of optimized locations selected for biomass supply counties (red bar), CSP facilities (green bar), and biorefinery (blue bar).

To better understand the variation of biomass supply chain configurations, each input parameter listed in Table 3.1 was divided into sub-ranges and plotted against decision variables including the number of supply counties and the number of CSP facilities. As indicated by Fig. 3.6a, the number of supply counties shows a consistent variation pattern with changes of cropland usage rate. The associated box plots shows a narrowing of range and lower median value with increased cropland usage rate. The number of supply counties is similarly affected by the change of biomass yield rate and conversion rate (Fig. 3.6b, e), but not by the change of logistics loss rate and raw biomass transportation cost (Fig. 3.6c, d). It is reasonable to draw the conclusion that high cropland usage rate, high biomass yield, and efficient conversion technology are more likely to reduce the number of supply counties. This conclusion makes sense since cropland usage rate and biomass yield rate are both related to the mount of biomass supply. When their values increase, sufficient biomass

supply can be secured and thus fewer biomass supply counties are required. Similarly, when biomass-ethanol conversion is more efficient, less biomass supply is required to meet the same bioethanol production demand.

Raw biomass transportation cost and logistics loss rate have direct impacts on the numbers of CSP facilities, where the interquartile range and median value have a consistent variation pattern. The increased raw biomass transportation cost and logistics loss rate would increase the overall cost of moving biomass from supply counties to CSP facilities (Fig. 3.7c, d). The optimal supply chain configurations would suggest building more CSP facilities near biomass supply counties to reduce the logistics loss and transportation costs. Figs. 3.7c and 3.7d show an increasing trend of variation of the number of CSP facilities, which indicates that the number of CSP facilities is more stable when logistics loss rate or raw biomass transportation cost is low. The changes of cropland usage rate, biomass yield rate, and biomass-ethanol conversion rate do not impose direct impacts on the number of CSP facilities given the fact that their mean values are not significantly changed (Fig. 3.7a, b, e). However, increasing these three inputs could highly affect the biomass sourcing area and subsequently increase the average transportation distance. This indicates that CSP facilities are more severely affected by the transportation cost than transportation distance.

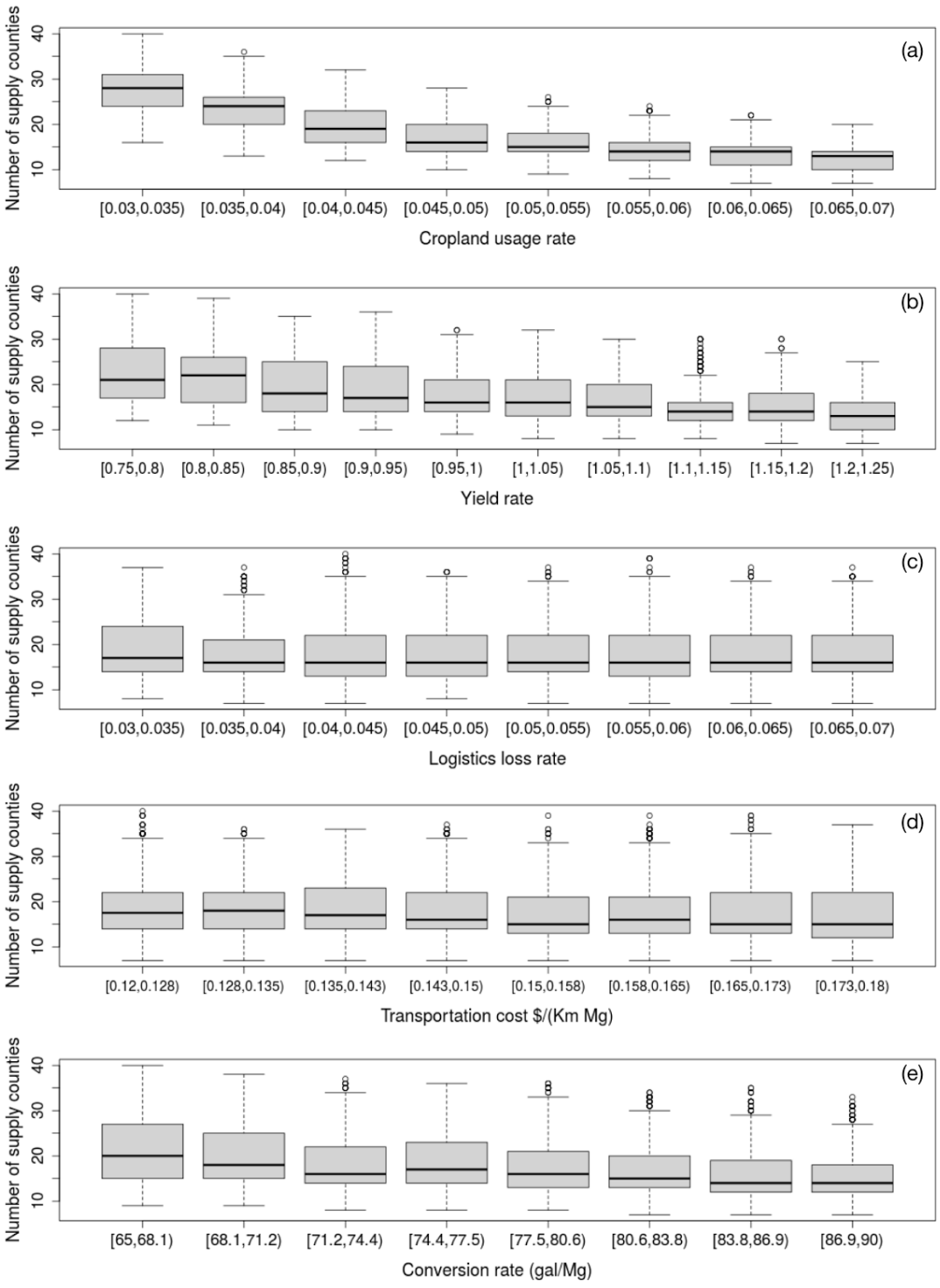


Figure 3.6. Boxplots of the probability distribution of the number of supply counties as related to (a) cropland usage rate, (b) biomass yield rate, (c) logistics loss rate, (d) raw biomass transportation cost, and (e) biomass-ethanol conversion rate. Boxes extend from 25th to 75th percentile, middle horizontal line within the box indicates the median projection, and whiskers are at maximum 1.5 interquartile range.

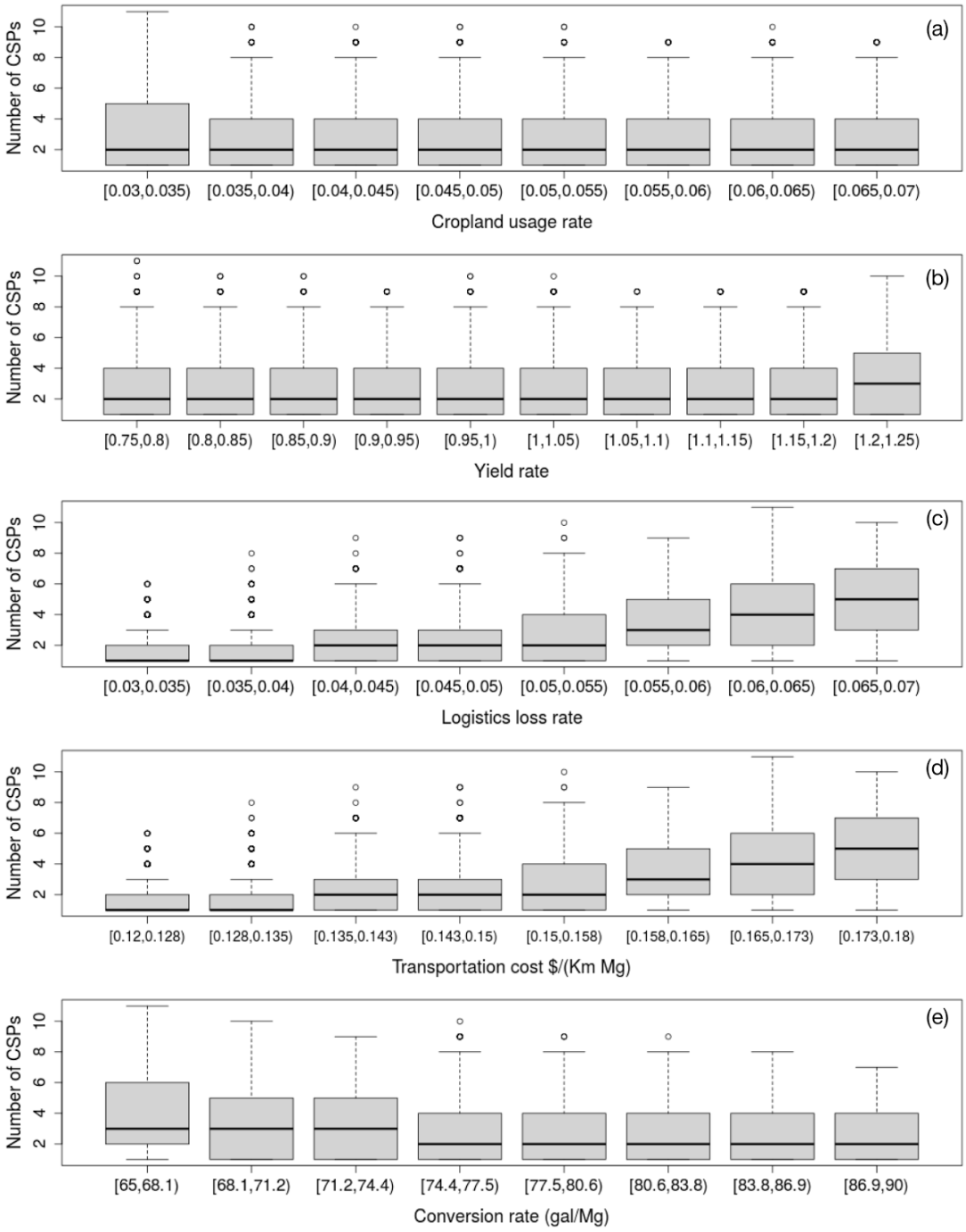


Figure 3.7. Boxplots of the probability distribution of the number of CSP facilities as related to (a) cropland usage rate, (b) biomass yield rate, (c) logistics loss rate, (d) raw biomass transportation cost, and (e) biomass-ethanol conversion rate. Boxes extend from 25th to 75th percentile, middle horizontal line within the box indicates the median projection, and whiskers are at maximum 1.5 interquartile range.

3.4.3 Local Sensitivity Analysis

Considering the CSP and biorefinery to be located at Jefferson County, which is the highest possible location of the 7,000 scenarios, local sensitivity analysis has been conducted to quantify the changes of each individual input parameter on the optimal ethanol production costs when others are kept at the same constant level. The results show that the changes of the biomass-ethanol conversion rate have the most significant impact on ethanol production costs, where a 15% increase of the conversion rate would reduce production costs by 18% (Figure 3.8). The increased conversion rate by technology improvement would reduce the demand of biomass feedstock, which results in a decrease in both biomass procurement and transportation costs. Increased cropland usage rate and biomass yield would also reduce ethanol production costs. Both increased cropland usage rate and yield directly increase biomass availability in each supply county, which reduces the required biomass sourcing area and associated biomass transportation costs. Both increased transportation cost and logistics losses would result in higher production costs, where the changes of transportation cost impose higher impact and the changes of logistics loss rate only showed a slight impact.

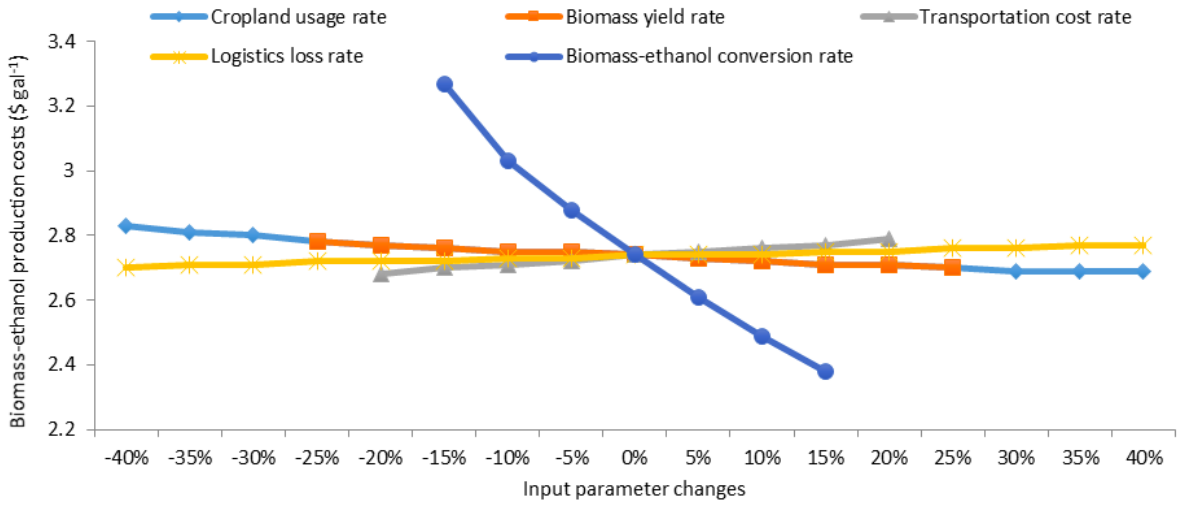


Figure 3.8 Local sensitivity analysis of the biomass supply chain optimization

3.4.4 Global Sensitivity Analysis

The results of local sensitivity analysis only quantify the effect of individual uncertain factor by changing one at a time, without considering the integrated effect when multiple factors change simultaneously. Hence, we further conducted global sensitivity analysis of these factors with regard to their entire parameter distributions. First-order and total effect sensitivity indices are calculated for yield rate, cropland usage rate, raw biomass transportation cost, logistics loss rate and conversion rate (Table 3.3 and Table 3.4), based on Eq 3.6 and Eq 3.7. The empirical confidence intervals (CI) were estimated using bootstrap approach. One hundred bootstrap replicas of the sample have been collected at a given parameter sample size (N=1,000), and 95% CI of these estimates were reported.

Table 3.3 First-order effect global sensitivity index with bootstrap confidence intervals

Parameter	Ethanol production cost			Number of CSPs			Number of supply counties		
	Rank	S_i	S_i with 95% CI	Rank	S_i	S_i with 95% CI	Rank	S_i	S_i with 95% CI
Yield rate	4	0.008	-0.04-0.088	4	0.004	-0.118-0.107	2	0.351	0.159-0.542
Cropland usage rate	2	0.027	0.003-0.125	5	0.000	-0.125-0.138	1	0.444	0.206-0.657
Raw biomass transportation cost	3	0.013	-0.021-0.197	1	0.285	0.077-0.456	4	0.097	-0.034-0.206
Logistics loss rate	5	0.006	-0.043-0.083	2	0.140	0.007-0.300	5	0.034	-0.04-0.119
Conversion rate	1	0.863	0.837-0.958	3	0.053	-0.101-0.173	3	0.258	0.104-0.440

Table 3.4 Total effect global sensitivity index with bootstrap confidence intervals

Parameter	Ethanol production cost			Number of CSPs			Number of supply counties		
	Rank	S_{Ti}	S_{Ti} with 95% CI	Rank	S_{Ti}	S_{Ti} with 95% CI	Rank	S_{Ti}	S_{Ti} with 95% CI
Yield rate	4	0.070	0.007-0.180	4	0.532	0.483-0.591	2	0.422	0.374-0.464
Cropland usage rate	2	0.096	0.023-0.230	5	0.504	0.440-0.565	1	0.502	0.432-0.545
Raw biomass transportation cost	3	0.085	0.010-0.195	1	0.829	0.753-0.913	4	0.124	0.432-0.545
Logistics loss rate	5	0.061	0.005-0.087	3	0.544	0.477-0.601	5	0.093	0.076-0.107
Conversion rate	1	0.946	0.733-0.985	2	0.585	0.509-0.645	3	0.326	0.291-0.359

The results show that biomass-ethanol conversion rate is the dominating input for ethanol production costs, which agrees with the local sensitivity analysis. The conversion rate, however, does not impose a significant impact on the selection of biomass supply counties and CSP facilities. Raw biomass transportation cost and the cropland usage rate are the two most important factors to the selection of CSPs and supply counties, respectively. Logistics loss rate would not affect much on production costs and the selection of biomass supply counties, but it has an influential impact on the selection of CSP facilities. To better illustrate the rank and magnitude of the first-order and total-effect sensitivity indices, two grid color plots are provided (Figure 3.9).

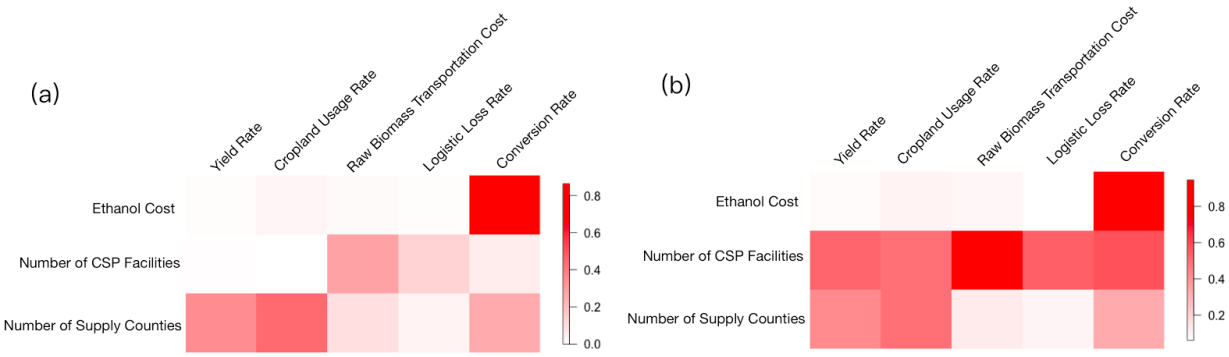


Figure 3.9 The results of global sensitivity analysis, first-order sensitivity indices (a) and total-effects sensitivity indices (b), of five input factors (*Yield Rate, Cropland Usage Rate, Raw biomass transportation cost, Logistic Loss Rate* and *Conversion Rate*) with respect to the biomass supply chain optimization objective – *Minimized Ethanol Production Costs* and two decision variables – *Number of CSP Facilities* and *Number of Supply Counties.*

By comparing the difference between first-order and total-effect indices, the interaction among inputs is found to be more obvious for the number of CSPs and number of supply counties. The impact of raw biomass transportation cost on facility location selection was increased from 0.285 of its first-order sensitivity index to 0.829 of its total-effect sensitivity index. This indicates that raw biomass transportation cost itself accounts for 28.5% of the variation of number of CSPs in the optimization model; but it accounts for 82.9% of the variation by considering the changes along with other inputs.

3.5 WHAT-IF SCENARIO ANALYSIS

The CyberGIS-BioScope decision support system is enabled by geodesign features to support analysis that is capable of responding to changes in the system. Given an established biomass supply chain in Illinois from the base case (Figure 3.5: Left), we performed what-if scenario analysis to help decision makers to evaluate decisions in case unexpected events occur in future

supply chain operations. Scenarios were evaluated by considering changes in biomass yield, raw biomass transportation cost, and biofuel demand.

3.5.1 Changes in Biomass Yield

In the management of biomass supply chains, insufficient biomass supply often occurs as the results of unexpected weather and pest disease. We assumed that the biomass yield in southern Illinois region is reduced by 15% due to weather or other factors. If we directly rerun the optimization model with the current available supply counties and biomass processing sites, no solution could be found due to the insufficient biomass supply to meet the demand. Thus, we allowed other candidate counties for biomass supply while maintaining the same demand. The results show that two counties (Effingham and Union) are selected as additional biomass suppliers (Figure 3.10). This change increases biomass transportation and procurement costs, because of a larger biomass sourcing area, and results in a \$0.01 gal^{-1} increase in ethanol production costs (Table 3.5).

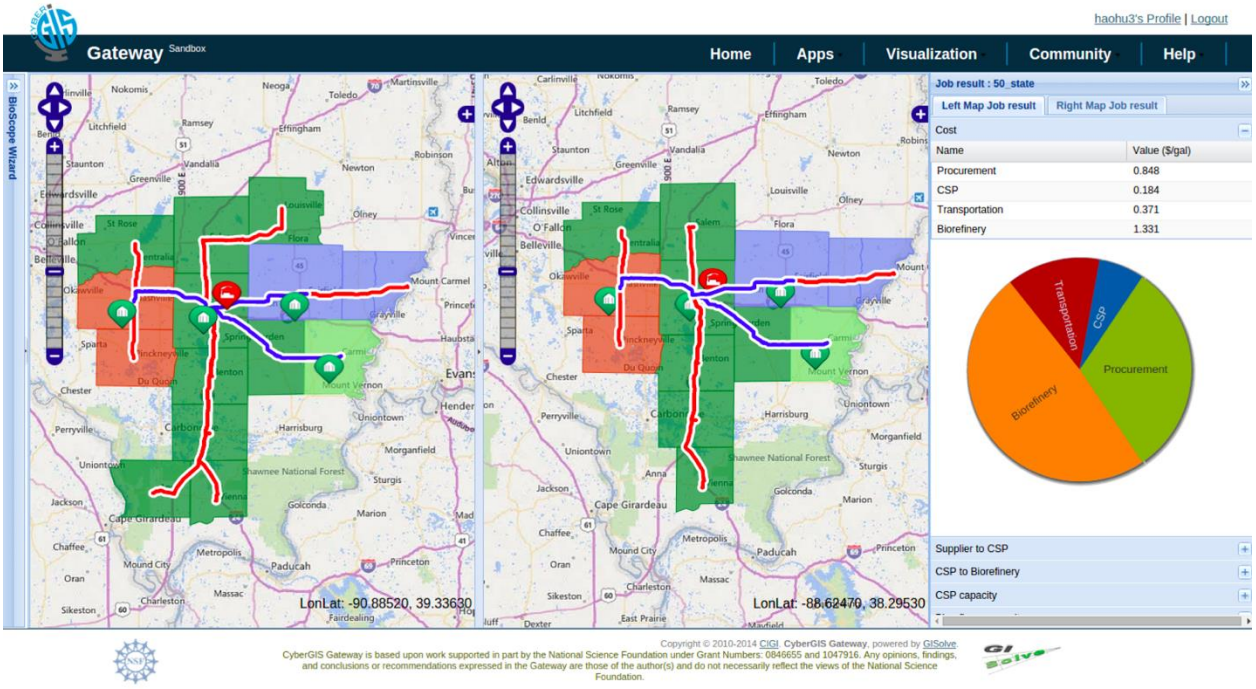


Figure 3.10 Biomass supply chain decisions in Illinois. Left: 15% decrease of biomass availability in southern Illinois. Right: base case.

3.5.2 Changes in Raw Biomass Transportation Cost

The change of raw biomass transportation cost could also change the optimal biomass supply chain configurations, especially considering the change of transportation cost for delivering raw biomass to CSP facilities. The solutions usually rely on how the optimization balances the trade-off between preprocessing biomass locally through building additional nearby CSP facilities or delivering them to a remote CSP facility with high raw biomass transportation cost. To evaluate this scenario, we assumed that the raw biomass transportation cost is increased by 20% compared to the base case. We regenerated the solutions in two cases: 1) with and 2) without building additional CSP facilities. The scenario without building additional CSPs generated the same pattern (Figure 3.11: Right) as that from the base case, but with an increase in the transportation cost from \$0.35 gal^{-1} to \$0.396 gal^{-1} (Table 5). The scenario with considering additional CSP

facilities generated a slightly different pattern by recommending a new CSP facility in Clinton County (Figure 3.11: Left). Although this decision leads to an increase of CSP cost from \$0.184 gal^{-1} to \$0.191 gal^{-1}, the transportation cost only increase from \$0.35 gal^{-1} to \$0.388 gal^{-1}, resulting in a more cost-efficient solution compared to the case without considering additional CSPs (Table 5).

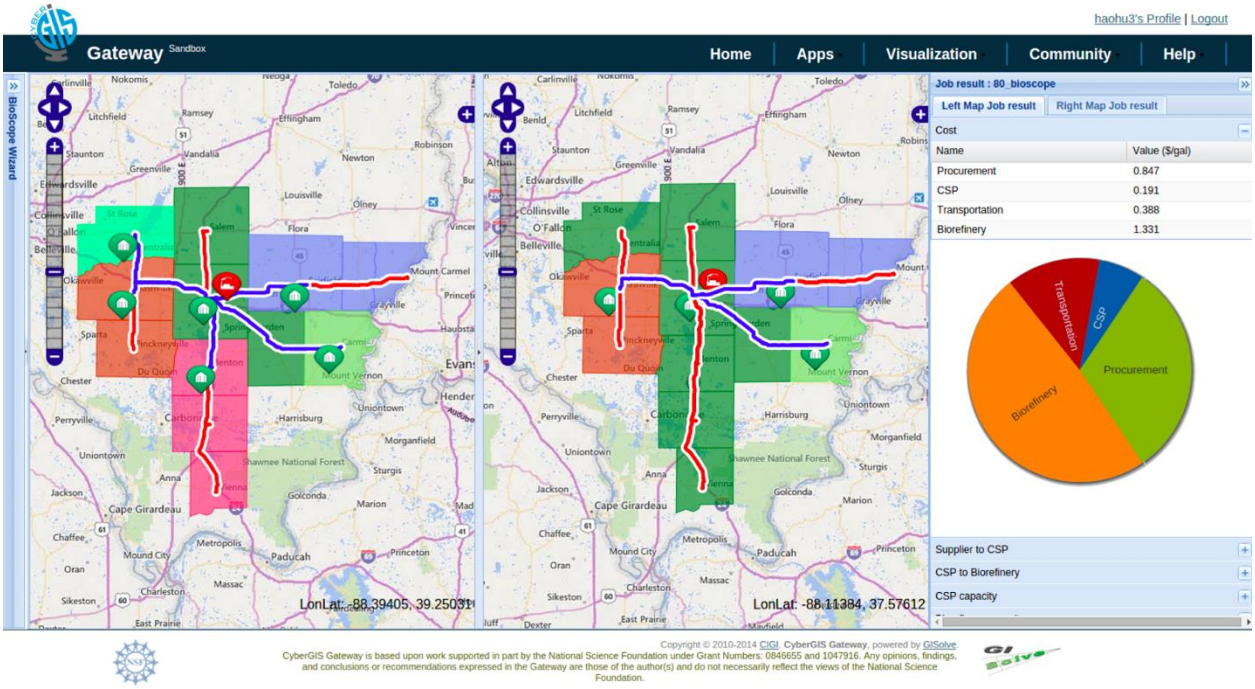


Figure 3.11. Biomass supply chain decisions in Illinois. Left: raw biomass transportation cost increases by 20% with considering new CSP facilities. Right: raw biomass transportation cost increases by 20% without considering new CSPs.

3.5.3 Changes in Biofuel Demand

As required by the renewable fuel standard, the demand for cellulosic biofuel will be increasing in the future. To demonstrate a representative case study, we assumed that annual ethanol demand increases from 160,000,000 gal in the base case to 200,000,000 gal. Meanwhile, considering potential government policies to incentivize more growth of bioenergy crops, we increased the cropland usage rate up to 6%. Similar to the previous case study, we designed two scenarios to

determine the location of supplemental biomass supply counties 1) with, and 2) without building additional CSP facilities.

Figure 3.12 shows the two decision results suggested by the cyberGIS system respectively. Both cases suggest adding supplemental biomass supplier only in Union County. Additional CSP facility in Clinton is selected when additional CSPs are considered. By further examining the ethanol production costs, we found that the case considering additional CSP facilities ($2.672 gal^{-1}) is more cost-efficient compared to the case without considering additional CSP facilities ($2.674 gal^{-1}) (Table 3.5). Meanwhile, the total ethanol production costs (in both cases) are less than that of the base case ($2.713 gal^{-1}), indicating that the ethanol production costs can be reduced with the increase in ethanol demand.

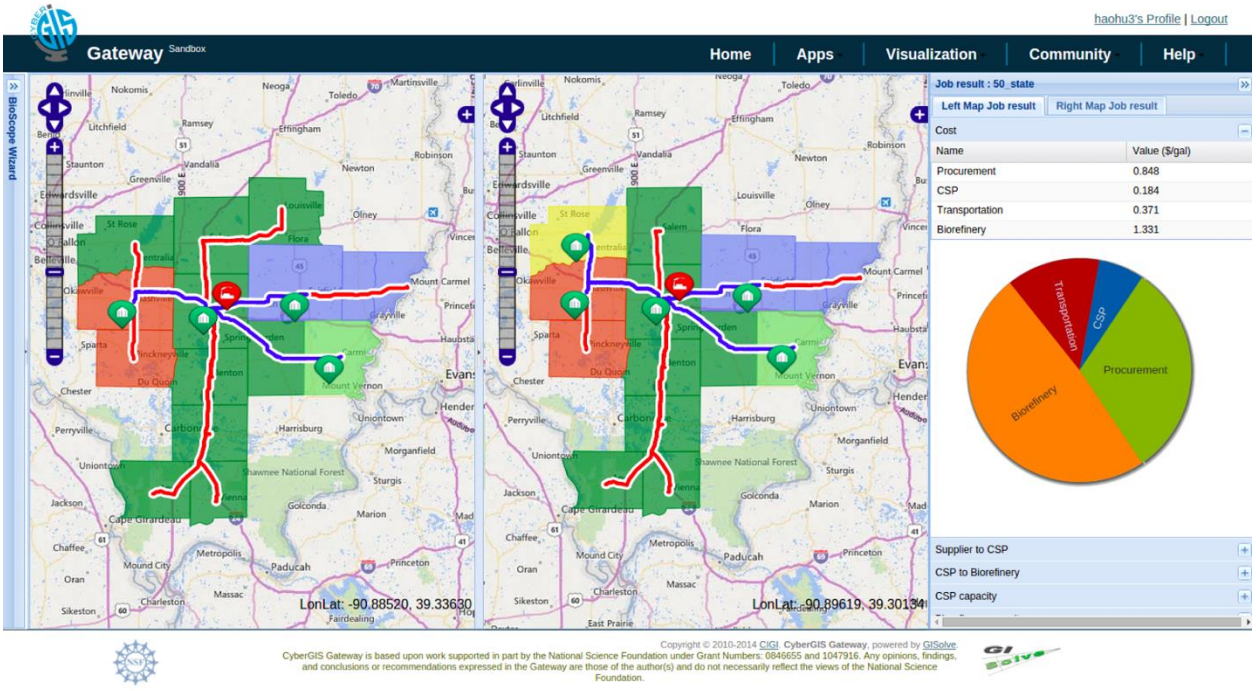


Figure 3.12 Biomass supply chain decisions in Illinois. Left: ethanol demand increases to 200,000,000 gal/y, cropland usage rate increases to 6%, but no additional CSP facilities are considered. Right: ethanol demand increases to 200,000,000 gal/y, cropland usage rate increases to 6%, with additional CSP facilities considered.

Table 3.5 Miscanthus supply chain cost summary in different scenarios (unit in $/gal)

Scenarios	Procurement Cost	Transportation Cost	CSP Cost	Biorefinery Cost	Total Cost
Base case	0.847	0.35	0.184	1.331	2.713
Reduced biomass yield	0.848	0.371	0.184	1.331	2.723
Increased raw biomass transportation cost	0.847	0.396	0.184	1.331	2.758
Increased raw biomass transportation cost, allowing additional CSPs	0.847	0.388	0.191	1.331	2.757
Increased demand	0.846	0.358	0.182	1.288	2.674
Increased demand, allowing additional CSPs	0.846	0.354	0.184	1.288	2.672

3.6 DISCUSSION

We started with the argument that uncertainty is associated with biomass supply chain optimization models, and that applying uncertainty and sensitivity analysis, based on Monte Carlo methods, leads to a better understanding of how and at what level various sources of uncertainty would impact on optimal ethanol production costs and supply chain configurations. The implementation of uncertainty and sensitivity analysis is powered by cyberGIS capabilities so that complex biomass supply chain models, computationally intensive uncertainty analysis, and visualization of model results are presented as an integrated spatial decision support system to decision makers. The results of uncertainty analysis are presented as distributions of model outputs, such as ethanol production costs and supply chain configurations, including the number of supply counties, storage facilities, or biorefineries (Section 3.4.2). Propagation of uncertainty can be quantified by basic measures of the resulting distributions such as range, shape, skew, and standard deviation.

The implementation of what-if scenario analysis aims to evaluate uncertainty from a different perspective. The results of what-if scenario analysis are more focused on individual decisions that compared to the base case. We considered three scenarios in which biomass yield, transportation cost, and biofuel demand are subject to change after the initial configurations of a biomass supply chain is implemented. In application, depending on the interest of decision makers, more scenarios can be included in what-if scenario analysis to explore the impact on cost-effective biomass supply chains.

Ultimately, our approach aims to provide decision support on where to make a change in a biomass supply chain given the presence of uncertainty in system inputs. For decision makers, two factors need to be considered when prioritizing where to make a change. First, we need to quantify how much influence the uncertain input has on the result. Second, we need to consider the ability to make a change in any given situation. Ideally we have the ability to make a change and the change will be influential. In some cases, we actually only have varying degrees of one or the other. For example, the results of global and local sensitivity analysis in our case study indicated that the most influential factor to ethanol production costs is biomass-ethanol conversion rate, followed by cropland usage rate, raw biomass transportation cost, biomass yield rate and logistics loss rate with the least (Table 3.3 and Table 3.4). By only considering the influence, most investments should go to technological inventions, which improve the biomass-ethanol conversion rate. However, improving the efficiency of biomass-ethanol conversion might be a long-term process. Therefore, investments on increasing the biomass yield, exploring more cropland for bioenergy crops, or reducing the logistics loss rate might be better strategies in the short term.

To further understand the impact of potential investments, the results of uncertainty and sensitivity analysis are extended to conduct a risk analysis on the optimal ethanol production costs. Three investment strategies are assumed to reduce ethanol production costs:

- Strategy 1: Improve the range of biomass-ethanol conversion rate from 66-90 Gal/Mg to 78-90 Gal/Mg.
- Strategy 2: Reduce the range of logistic loss rate from 3%-7% to 3%-4%.
- Strategy 3: Improve the range of biomass yield rate from 0.75-1.25 to 1.2-1.25 and the range of cropland usage rate from 3%-7% to 6%-7%.

Strategy 1 represents the plan to invest on improving the biomass-ethanol conversion rate. Investment in reducing the logistics loss rate, which is the least influential parameter reported by sensitivity analysis, is considered in Strategy 2. In Strategy 3, we assumed investments are spent on improving biomass yield and cropland usage rate. There is a huge potential in genetics and breeding to improve the yield of bioenergy crops, especially for Miscanthus which has not been studied extensively. Meanwhile, potential incentives from government policies may increase cropland usage rate so that more agricultural land can be converted for growing bioenergy crops. We selected these three strategies because they are representative to show the impact of different investment strategies on the distribution of optimal ethanol production costs.

For the purpose of comparison, we used density curves to examine the distributions of ethanol production costs (Figure 3.13). The red density curve indicates the original distribution of ethanol production costs, which is directly converted from the histogram in Figure 2.4a. By comparison, it is obvious that higher biomass-ethanol conversion rate (green) will significantly reduce the mean and variance of ethanol production costs. If the ability to improve the conversion rate is limited, increasing the cropland usage rate and the biomass yield rate (purple) appears to be a better

investment strategy than reducing the logistics loss rate (blue) given their similar variance but smaller mean value in the former. Oftentimes, the ability to make a change is known to decision makers. The information they need to reduce the risk is whether the change will be influential and the robustness of the influence. In application, even if you could create conditions favoring a certain strategy, it may not be that influential to your final outcome—and therefore may not be a good strategy.

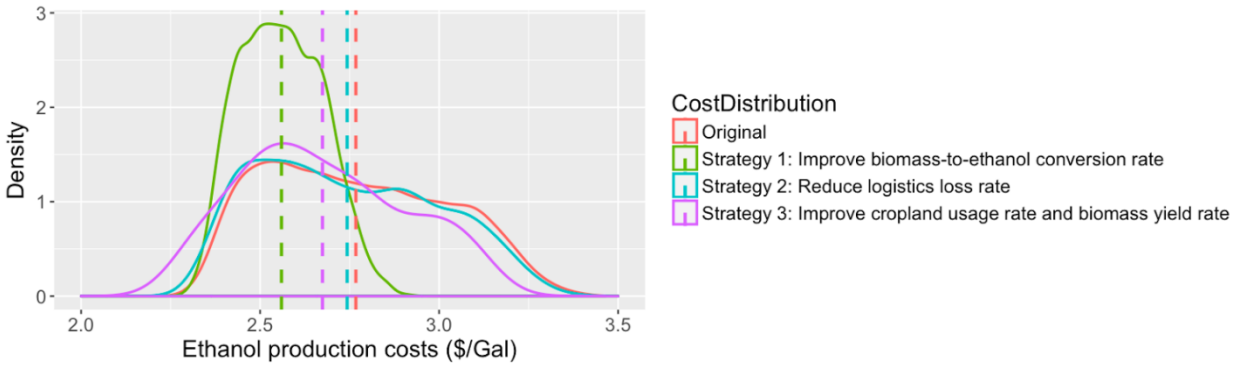


Figure 3.13 –Distribution of optimized bioethanol production cost. Dashed line represents the mean of each distribution.

3.7 CONCLUSION AND FUTURE WORK

In the biomass supply chain design and operational planning, model-based approaches are sometimes unreliable when most of the data are fraught with uncertainties. Uncertainty and sensitivity analysis techniques offer an accessible treatment via the quantification of uncertainty propagations and sensitivity measurement. This paper describes a cyberGIS-enabled decision support system that provides innovative analytical and computational capabilities for optimizing biomass supply chains under uncertainties. The system serves as a new integrated approach to data management, mathematical modeling, uncertainty and sensitivity analysis, what-if scenario analysis, and result representation and visualization for the biomass supply chain optimization. Leveraging high-performance computing capabilities provisioned by advanced cyberinfrastructure,

this cyberGIS approach enables Monte Carlo based uncertainty and sensitivity analysis for optimization modeling by resolving significant computational intensity. Furthermore, the approach supports dynamic management decisions through what-if scenario analysis responding to uncertain situations in supply chain operations.

In the case study of optimizing Miscanthus supply chain in Illinois, United States, a Monte Carlo based optimization modeling analysis with 7,000 scenarios has been conducted to quantify the uncertainty and sensitivity impact of various input factors on ethanol production costs and optimal biomass supply chain configurations. The results from uncertainty analysis showed that the minimal ethanol production costs range from \$2.30 to \$3.43 gal^{-1}. The mean production costs are \$2.77 gal^{-1} with a standard deviation of \$0.25 gal^{-1}. The sensitivity analysis showed that biomass-ethanol conversion rate is the most influential factor to ethanol production costs as expected, which indicates that the development of conversion technology is critical in reducing cellulosic ethanol production costs. The results also showed that the optimal biomass supply chain infrastructure is sensitive to changes in biomass yield, transportation cost, and logistics loss rate. For the same system demand, higher county-level biomass availability would require fewer supply sites, whereas higher transportation cost and logistics loss rate would result in more centralized storage and preprocessing facilities.

The proposed cyberGIS approach bridges the gaps between research, development, and implementation of biomass supply chain optimization under uncertainties. This approach generalizes and streamlines data management, computation, and visualization components, so it is expected to work on biomass supply chain optimization applications customized by different model and data inputs. Future work will seek to further understand the complexity and sustainability of biomass-biofuel supply systems. Multi-objective optimization with economic,

environmental, and social measurements is desirable to be incorporated into the decision support system. Another opportunity is to combine models that are at finer scale of biomass supply chain management and implement operational level decision making. We plan to incorporate optimal control of such processes as harvest scheduling, inventory planning, and transportation management into the current cyberGIS-based decision support system.

CHAPTER 4

BIOMASS SUPPLY CHAIN OPTIMIZATION WITH UNCERTAIN BIOMASS AVAILABILITY

This chapter first describes a Bayesian Hierarchical Model (BHM) to quantify the biomass supply uncertainty caused by weather dynamics and then proposes a stochastic programming model to account for such uncertainty in the optimization of biomass feedstock provision. Compared to alternative statistical models that address the relationship between biomass yield and weather, the BHM model captures the variation of biomass yields that can be explained by weather change with the consideration of spatiotemporal non-stationarity. The results of biomass yield analysis are then used for scenario constructions in a stochastic programming model to provide decision support for strategic level biomass supply chain development considering the uncertainties of biomass supply.

This chapter cannot be completed without successful team support from Drs. Tao Lin, Bo Li, and Shaowen Wang. Mr. Hu led the overall research design, methodology development, case study analysis, and manuscript writing. Dr. Lin participated in the overall research design and provided insightful suggestions for optimization modeling part. Dr. Li contributed particularly to the statistical modeling of Bayesian Hierarchical methods. Dr. Wang participated in the overall research design and results discussions, and led the draft revision. The content of this chapter is based on a manuscript submitted to *The Annals of the American Association of Geographers*, with the expansion of integration with stochastic optimization method, updated case study design and result discussions.

Abstract. *Spatial optimization is an interdisciplinary field that integrates spatial data analysis and modeling, operation research, and domain specific knowledge to enable geospatial problem*

solving and decision-making. In the real world, spatial optimization is subject to uncertainty given the nature of spatiotemporal data and associated modeling approaches. Stochastic programming (SP) is a widely adopted approach to account for uncertain factors in optimization modeling. However, spatiotemporal characteristics of uncertainty are not well understood while developing SP approaches.

In this study, we present an integrated approach that captures spatiotemporal data uncertainty in a supply chain optimization problem. Based on hierarchical Bayesian modeling that accounts for varying spatial and temporal relationships between outcomes and covariates, this novel approach captures spatiotemporally explicit uncertainty associated within the data as opposed to the more traditional approach without considering spatial and/or temporal relationships when constructing discrete scenarios analysis in SP.

We focus on a bioenergy supply chain problem to optimize supply chain infrastructure configurations, biomass feedstock (cornstover) provision and logistics operations. Corn yield, largely impacted by weather dynamics, are considered as a major source of uncertainty in the optimization of biomass supply chain.

4.1 INTRODUCTION

Biomass supply chains are complex systems that consist of four major subsystems: production, processing and manufacturing, distribution, and utilization (Awudu and Zhang, 2012). Biomass production subsystem is spatially and temporally explicit and is sensitive to changes in climate and weather patterns. Fluctuations in agricultural productivity impose significant challenges on the design and operations of integrated biomass supply chain systems. As a result, a key challenge

for biomass supply chain modeling is to incorporate uncertainties of agricultural production across space and time to support planning and management decisions.

As a major uncertainty of biomass supply chain, biomass yield varies annually as a result of weather dynamics for both agricultural residuals and perennial energy crops. The temporal changes of biomass yield would significantly affect the optimal supply chain configurations, especially considering a life cycle of biorefinery facility with more than 10 years. Therefore, understanding the effects of weather variability on biomass yields is central to design robust biomass supply chain management strategies (Lobell and Burke 2010; Olesen *et al.* 2011).

Two main modeling approaches have been extensively studied to quantitatively understand the relationship between biomass yields and weather dynamics, namely process-based models and statistical models. Process-based models are also referred to as biomass/crop simulation models. In addition to weather variables, this type of models, for example, DSSAT (Jones *et al.* 2003), often require various input data such as cultivar, soil conditions, and farm management that are unavailable or expensive to obtain for biomass yield studies at large geographic scales. Alternatively, statistical models have been developed and tested to quantify the relationship between biomass yields and weather (Schlenker and Roberts 2009; Lobell and Burke 2010; Bornn and Zidek 2012; Ray *et al.* 2015; McGrath *et al.* 2015). However, two major issues are often discussed in existing regression methods for biomass yield estimations. First, the impact of weather on biomass yields might follow a spatially non-stationary process, i.e., regression coefficients do not necessarily remain fixed from location to location, especially when a study area covers a variety of spatially heterogeneous landscapes, soil properties or agronomic practices (Sharma *et al.* 2011; Cai *et al.* 2014). Second, the error terms are spatially correlated, which violates the assumption of independent and normally distributed residuals in linear regression models. This

spatial dependence of residuals may indicate that the model developed is inadequate to fully explain the data, thus may result in poor model fitting and less accurate predictions (Hoeting 2009; Jiang *et al.* 2009; Bornn and Zidek 2012).

One of the major reasons for spatially non-stationary processes can be explained by the fact that relationships between the outcomes and covariates are intrinsically different across space (Fortheringham *et al.* 1998). For example, the impact of precipitation on biomass yield is perhaps much stronger for rain-fed regions compared to irrigated regions. Similarly, surface characteristics that are varying over space such as soil moisture could respond to precipitation differently even with the same weather conditions, which may result in different effects on biomass yield. Therefore, the stationary coefficients (i.e. intercept and slopes) of the classical multiple linear regression models present a challenge in describing the relationships between biomass yields and weather factors on large spatial scales.

Three major approaches for modeling spatially non-stationary processes are commonly used in past studies: geographically weighted regression (GWR) model (Brunsdon *et al.* 1996; Fotheringham *et al.* 1998; Brunsdon *et al.* 2001; Fotheringham *et al.* 2002), Bayesian spatially-varying coefficients (SVC) model (Gelfand *et al.* 2003; Banerjee *et al.* 2014), and Moran's eigenvector-based spatial regression approach which is often referred to as eigenvector spatial filtering (ESF: Griffith, 2008; Murakami et al. 2017). GWR, SVC and ESF are all capable of estimating spatially varying associations between outcomes and covariates but their implementations are different. In GWR, the estimated coefficient surface varies from location to location with smoothness determined by a kernel function and bandwidth. In SVC, the overall mean of coefficients is estimated first and then local deviations from the mean are estimated by applying a spatial random effect such as conditional autoregressive (CAR) models (Waller *et al.*

2007). Because prior information such as the model for spatial random effects can be included in the model, SVC models fit nicely into the Bayesian hierarchical spatial modeling framework. In terms of handling the spatial dependency or similarity between neighbors inherent in data, GWR attempts to capture most similarity using spatially varying coefficients and then simply assume errors to be independent (Fotheringham *et al.* 2002), whereas SVC directly models the autocorrelation by decomposing the residuals into structured random effects and white noise (Waller *et al.* 2007). Therefore, both GWR and SVC mitigate the issue of spatial autocorrelation without introducing common spatial regression techniques such as spatial lag and error models (Anselin 2013), though the degree of mitigation is limited since that these two methods are initially developed for different purposes (Basile *et al.* 2014). Unlike GWR and SVC, the ESF-based approach allows for controlling the number of parameters (i.e., model complexity) through variable selection. However, both Helbich and Griffith (2016) and Murakami et al. (2017) demonstrated the instability of the ESF-based approach where it can suffer just as basic GWR does with respect to limitations of multicollinearity between the local parameter estimates and the assumption of same degree of spatial smoothness. Therefore, ESF-based methods and derived methods with temporal dimension are not further discussed in the context of this thesis.

For both GWR and SVC models, a new problem arises when spatiotemporal data (also referred to as panel data in literature) are used – the local relationship between outcomes and covariates may not be time-invariant (Gelfand *et al.* 2003; Choi *et al.* 2012). When applying to regression modeling of crop yields, it means that the associations between crop yields and weather may be subject to change over space and time. Therefore, there is a need for appropriate space-time statistical models to be developed (An *et al.* 2015).

Research efforts on modeling the spatially or spatiotemporally varying association between crop yields and weather factors have been pursued with GWR models. Sharma *et al.* (2011) employed the GWR model to account for the impact of the spatial non-stationarity relationship between crop yields and precipitation for ninety-three counties in Nebraska. Their results showed that the performance of GWR in estimating yield for both corn and soybean under irrigated and rain-fed conditions was significantly better than the performance of ordinary least square (OLS) models. Cai *et al.* (2014) expanded the idea of GWR to geographically weighted panel regression (GWPR) to study crop yield response to weather variations for 958 US counties. According to their results, precipitation effects are sensitive to the existence of irrigation systems. Although the spatial non-stationary relationship between crop yields and weather are considered in their GWPR model, they assumed that this relationship remains fixed over time, which may need further investigation.

Many biofuels supply chain optimization studies have been conducted to minimize production costs or maximize profits using deterministic modeling (Rentizelas et al., 2009; Huang et al., 2010; You and Wang, 2011; Lin et al., 2013; 2014). Considering the changes of biomass supply, demand, and fuel prices, uncertainty and sensitivity analyses have been conducted using deterministic optimization models to quantify the impact of input parameters on optimal bioenergy supply chain (Kim et al., 2011; Hu et al., 2017). Recently, several stochastic optimization models have been developed to quantify the impact of uncertainties on optimal design and operations of biomass supply chain (Dal-Mas et al., 2011; Chen and Fan; 2012; Awudu and Zhang, 2012; Tong et al., 2013). Most studies aim for single year optimization, assuming the constant or average input parameters such as biomass supply and biofuel demands. Annual changes of biomass supply in a long-term production life cycle of biorefinery operation have not been well studied and this would affect strategic planning decisions such as optimal facility capacity and location. The long-term

historical pattern of biomass yield with spatial specific information is critical to quantify spatial and temporal changes in biomass supplies.

This study aims to develop a two-stage stochastic optimization model to provide decision support for strategic level biomass supply chain development considering the uncertainties of biomass supply in a long planning period. Particularly, a novel spatiotemporally varying coefficients (STVC) model is proposed to account for non-stationary responses of biomass yields to weather variables for the purpose of constructing scenarios in the stochastic programming model formulation. The STVC model examines the spatiotemporal varying relationship between corn yields and weather using county-level data in the Midwestern US for the period of 1981 to 2015. The rest of the chapter is organized as follows. First, we introduce methodological foundations for relevant models, including the proposed STVC model and the development of the two-stage stochastic programming model in details. Next, we combine two models to study the variability of corn yields in response to weather covariates and quantify the impact of uncertainties of biomass supply on optimal biomass supply chain configurations. The results of the model comparison are discussed accordingly. Finally, we conclude by summarizing our research findings and future work.

4.2 METHODLOGY

4.2.1 Geographical and Temporal Weighted Regression (GTWR) Model

Starting from the basic ordinary least squares (OLS) regression model for spatial data, the dependent variable is modeled as a linear function of a set of independent variables plus errors,

$$y_{(s)} = \mathbf{x}_{(s)}^T \boldsymbol{\beta} + \varepsilon_{(s)}, \tag{4.1}$$

where $y_{(s)}$ represents the outcome observed at location *s*, $\mathbf{x}_{(s)}$ is a set of *p* covariates including a column of 1's for the intercept, $\boldsymbol{\beta}$ is a $p \times 1$ vector of coefficients which can be estimated by OLS method, and $\varepsilon_{(s)}$ is white noise error. In the OLS model, the residuals are assumed to be independent and normally distributed, and represent vertical distances between the actual values of the dependent variable and their mean values. The estimates for the coefficients are chosen to minimize the sum of squared residuals (SSR). However, when OLS is applied to spatial data, the residuals are often correlated rather than being independent, thus more advanced methods are required to deal with the autocorrelation in the residuals.

When spatial data are temporally referenced, we have spatiotemporal data to be considered in regression analysis. Another term – panel data – is often used to describe such data in econometrics. Regression models developed for panel data are referred to as panel regression models. Compared to OLS, a panel regression model is capable of capturing the uniqueness of spatial effects. For example, a basic panel regression that incorporates spatial heterogeneity can be represented as:

$$y_{(s,t)} = \mathbf{x}_{(s,t)}^T \boldsymbol{\beta} + \alpha_{(s)} + \varepsilon_{(s,t)}, \tag{4.2}$$

where $y_{(s,t)}$ represents the dependent variable at location *s* in time *t*, $\mathbf{x}_{(s,t)}$ is a set of covariates specific to location and time, $\alpha_{(s)}$ is a site-specific term for controlling time-invariant spatial heterogeneity. There are two distinct approaches to modeling this site-specific term (Hsiao 2014). One is to treat $\alpha_{(s)}$ as a fixed but unknown parameter to estimate. In this case, equation (4.2) is known as a fixed effects model. Another approach is to treat $\alpha_{(s)}$ as drawn from an unknown population and thus random variables. In this case, equation (4.2) is known as a random effects model.

Although a spatial term is introduced in the panel regression model such as in Equation (4.2) to capture spatial heterogeneity, the constant coefficient vector $\boldsymbol{\beta}$ indicates a stationary relationship between the independent and dependent variables from location to location. However, for some applications this stationary assumption may not be appropriate. The idea of geographically weighted regression (GWR) is to deal with the situations where the impact of covariates on the outcome varies over spatial locations (Fotheringham *et al.* 1998; Brunsdon *et al.* 2001; Fotheringham *et al.* 2002). Compared to OLS and panel regression, GWR allows for spatially varying coefficients as in the following equation:

$$y_{(s)} = \mathbf{x}_{(s)}^T \boldsymbol{\beta}_{(s)} + \varepsilon_{(s)} \tag{4.3}$$

where $\boldsymbol{\beta}_{(s)}$ is now a $p \times 1$ vector of coefficients at location *s*, and it can be estimated by:

$$\hat{\boldsymbol{\beta}}_{(s)} = [\mathbf{X}^T \mathbf{W}(s) \mathbf{X}]^{-1} \mathbf{X}^T \mathbf{W}(s) \mathbf{Y} \tag{4.4}$$

where $\mathbf{Y}$ is the $n \times 1$ vector of dependent variable values; $\mathbf{X}$ is an $n \times p$ matrix of covariates in the form of $[\mathbf{x}_{(1)}^T; \mathbf{x}_{(1)}^T, ..., \mathbf{x}_{(n)}^T]^T$; and $\mathbf{W}(s)$ is an $n \times n$ diagonal weight matrix. $\mathbf{W}(s)$ can be either a discrete weight matrix where each entry is set to be one if its corresponding region is near location *s* and zero otherwise, or a continuous weight matrix based on a distance-decay function which places more weight on observations that are closer to location *s*.

When dealing with data that are both spatially and temporally referenced, GWR can be further expanded to geographical and temporal weighted regression (GTWR). The major difference lies in the definition of weight matrix which should capture both spatial and temporal effects from observations nearby in either space or time (Huang *et al.* 2010; Wu *et al*. 2014; Fotheringham *et al.* 2015). With spatial and temporal non-stationary effects combined, the new coefficients at location *s* and time *t* can be estimated by:

$$\hat{\boldsymbol{\beta}}_{(s,t)} = [\mathbf{X}^T \mathbf{W}(s,t)\mathbf{X}]^{-1} \mathbf{X}^T \mathbf{W}(s,t)\mathbf{Y} \tag{4.5}$$

where $\mathbf{W}(s,t) = diag(\alpha_1, \alpha_1, ..., \alpha_n)$ is the and *n* is the number of observations. Here the diagonal elements are space-time distance decay function and therefore, defining and measuring the so-called closeness in a space–time coordinate system is a key problem in the GTWR model.

4.2.2 Bayesian Spatiotemporally Varying Coefficients (STVC) Model

Similar as the GWR model defined in Equation (4.3), Gelfand *et al.* (2003) developed a spatially varying coefficient model as follows:

$$y_{(s)} = \mathbf{x}_{(s)}^T \boldsymbol{\beta} + \mathbf{x}_{(s)}^T \boldsymbol{\beta}_{(s)} + \varepsilon_{(s)} \tag{4.6}$$

where $\boldsymbol{\beta}_{(s)}$ is a second-order stationary mean zero Gaussian process independent of the white noise error process. Note that if we combine $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_{(s)}$, then Equation (4.6) will be exactly the same as Equation (4.3) for the GWR model. However, the formulation in (4.6) allows decomposing the total effects of covariates into an overall mean effect at global level, denoted by $\boldsymbol{\beta}$, and a local deviation at location *s* to the overall mean effect, denoted by $\boldsymbol{\beta}_{(s)}$. A popular choice for modeling the spatial random process $\boldsymbol{\beta}_{(s)}$ (Waller *et al.* 2007; Ozaki *et al.* 2008; Dong *et al.* 2016) is an intrinsic conditional autoregressive (ICAR) model. ICAR models have been widely used for the data that are aggregated over an area, because they allow the statistical inference at a particular region to borrow strength from adjacent and nearby regions. Specifically, the model for $\boldsymbol{\beta}_{(s)}$ governed by ICAR is given as:

$$\beta_{i(s)} \mid \beta_{i(s*)} \sim N\left(\frac{1}{m_s}\sum \beta_{i(s*)}, \frac{\sigma_{\eta_s}^2}{m_s}\right), \quad i = 1,2,\ldots, p \tag{4.7}$$

where $\beta_{i(s)}$ is the *i*th element in $\boldsymbol{\beta}_{(s)}$, s^* is the set of neighbors of region *s*, m_s is the cardinality of the set s^* or the number of neighbors, and $\sigma^2_{\eta_s}$ controls the magnitude of spatial variation. A constraint of $\sum_s \beta_{i(s)} = 0$ is usually applied to ensure the identifiability of the parameters. This structure is essentially a multivariate Gaussian distribution with a particular covariance matrix.

The SVC is typically constructed in a Bayesian hierarchical modeling (BHM) framework with Equations (4.6) and (4.7) constituting the first and the second level of the BHM. Given priors for all the unknown parameters in (4.6) and (4.7) the hierarchy is closed. Then Markov chain Monte Carlo (MCMC) algorithms are used to perform parameter estimation and draw statistical inference. Bayesian formulations of ICAR were discussed with details by Besag *et al.* (1991) and Besag and Kooperberg (1995). A nice feature of the BHM is that we can easily estimate the uncertainty along with the point estimates.

We extend the SVC model in Gelfand *et al.* (2003) for spatiotemporal data. The innovation of this extension is to allow the spatially varying coefficient to change over time. Accordingly, our model is expressed as:

$$y_{(s,t)} = \mathbf{x}^T_{(s,t)}\boldsymbol{\beta} + \mathbf{x}^T_{(s,t)}\boldsymbol{\beta}_{(s,t)} + \varepsilon_{(s,t)} \tag{4.8}$$

where each element in $\boldsymbol{\beta}_{(s,t)}$ is further decomposed as:

$$\boldsymbol{\beta}_{(s,t)} = \boldsymbol{\beta}_{(s)} + \boldsymbol{\alpha}_{(t)} \tag{4.9}$$

where $\boldsymbol{\beta}_{(s)}$ represents the spatial random effects that can take the form defined in Equation (4.7), and $\boldsymbol{\alpha}_{(t)}$ represents the temporal random effects that can be modelled by a random walk or an autoregressive process. For example, elements in term $\boldsymbol{\alpha}_{(t)}$ can be defined as the random walk process:

$$\alpha_{i(t)} = \alpha_{i(t-1)} + \varepsilon_{i(t)}, \quad i = 1,2,3,\ldots, p \tag{4.10}$$

where $\varepsilon_{i(t)}$ is white noise, and $\alpha_{i(0)} = 0$. The explanation behind a random walk process is that the current value of the random variable is determined by the past value plus an independent error term. As pointed out by Gelfand *et al.* (2003), Equation (4.9) is not the only way to decompose the spatiotemporally varying coefficients. More sophisticated forms of nested spatiotemporal relationships could be specified. In this study, we will focus on the process defined in Equation (4.9) where the spatial non-stationarity is captured by the ICAR model defined in Equation (4.7) and the temporal non-stationarity is captured by the random walk defined in Equation (4.10).

4.2.3 Two-stage Stochastic Biomass Supply Chain Optimization Model

The two-stage stochastic programming biomass supply chain optimization model is developed based on the BioScope model proposed by Lin et al. in (2013) by considering biomass supply variations. The stochastic BioScope model is to minimize the expected overall costs of an integrated biomass supply chain system over the entire planning period considering different biomass supply scenarios. The decisions that are optimized by the model include: 1) optimal numbers, locations, and capacities of centralized storage and preprocessing (CSP) and biorefinery facilities; 2) optimal purchase amount of biomass supply from each supplier each year; 3) optimal biomass transportation flow patterns each year.

The stochastic BioScope model is a mixed integer linear programming (MILP) model that was developed and solved using the Gurobi (http://www.gurobi.com/) optimizer. A list of set names, decision variables, and parameters used in the model is provided in "Nomenclature" in Appendix B. In the stochastic BioScope model, all the decisions related to long-term supply chain configurations are considered the first-stage decision variables that cannot be changed throughout the planning period for any scenario. They include the numbers, locations, and capacities of CSP

and biorefinery facilities. Biomass yield varies each year, and changes in biomass yield would affect biomass availability and procurement costs. Given the changes in biomass availability and costs, biomass procurement and associated transportation patterns are the second-stage decision variables. They can vary among different years in different scenarios.

For this stochastic optimization model, the overall expected biomass-ethanol production costs (Z) are comprised of four costs: biomass purchase costs (C_B^s), transportation related costs (C_M^s), CSP related costs (C_S), and biorefinery related costs (C_E) (Eq. 4.11). CSP and biorefinery related costs are related to first-level decision variables that would not be changed in different scenarios, while biomass purchase and transportation costs are related to second-level decision variables that would change in different scenarios with its expected probability density (ρ^s).

$$Minimize\ Z = \sum_s \rho^s \times (C_B^s + C_M^s) + C_S\ + C_E \tag{4.11}$$

Biomass supply

Biomass purchase costs (C_B^s) are changing in different scenarios based on the variation of biomass availability (Lin et al., 2017 manuscript). They are a function of the optimal biomass flow pattern ($f^{i,j,s}$) from supply sites to CSP sites and the county-level biomass production costs at the supply site ($c^{i,s}$) (Eq. 4.12). The total amount of biomass output from a biomass supply site should not exceed its biomass availability in scenario s (Eq. 4.13). County-level biomass production costs ($c^{i,s}$) and biomass availability ($b^{i,s}$) are two inputs related to biomass yield in scenario s. It is important to optimize the supply site selection as well as the quantity of biomass to purchase from each site.

$$C_B^s = \sum_i \sum_j c^{i,s} \times f^{i,j,s} \tag{4.12}$$

$$\sum_j f^{i,j,s} \leq b^{i,s} \tag{4.13}$$

Biomass transportation

Biomass transportation costs (C_M^s) are composed of variable ($C_{M_v}^s$) and fixed transportation costs ($C_{M_f}^s$) in scenario s (Eq. 4.14). The decision variables related to total biomass purchase costs are the amount of biomass flow from supply sites to CSP sites ($f^{i,j,s}$) and the amount of preprocessed biomass flow from CSP sites to biorefineries ($f^{j,k,s}$) in scenario s. Variable transportation costs are a function of the unit variable transportation cost (t_{v1}, t_{v2}), amount of biomass being transported ($f^{i,j,s}$, $f^{j,k,s}$), and the transportation distance ($d^{i,j}, d^{j,k}$) (Eq. 4.15). Fixed transportation costs that include loading and unloading costs depend on the unit fixed transportation cost(t_{f1}, t_{f2}) and the amount of biomass being transported ($f^{i,j,s}$, $f^{j,k,s}$) (Eq. 4.16). The shortest distances between the facilities within the system ($d^{i,j}$, $d^{j,k}$) are inputs calculated via OSRM (Luxen and Vetter, 2011) using the OpenStreetMap road network.

$$C_M^s = C_{M_v}^s + C_{M_f}^s \tag{4.14}$$

$$C_{M_v}^s = \sum_i \sum_j (t_{v1} \times f^{i,j,s} \times d^{i,j}) + \sum_j \sum_k (t_{v2} \times f^{j,k,s} \times d^{j,k}) \tag{4.15}$$

$$C_{M_f}^s = \sum_i \sum_j (t_{f1} \times f^{i,j,s}) + \sum_j \sum_k t_{f2} \times f^{j,k,s} \tag{4.16}$$

Centralized storage and preprocessing (CSP)

The costs related to CSP facilities (C_S) are composed of annual operating costs (C_{S_o}) and annual capital related costs (C_{S_c}) (Eq. 4.17). In this study, it is assumed that CSP facilities with different capacities incur the same unit operating costs (s_{op}). Therefore, annual operating costs are linearly dependent on the demand of biomass for CSP facilities (Eq. 4.18).

Annual capital costs are linearly dependent on the capital investment costs where a factor α (13.7%) is used to represent its relationship. To improve the accuracy, the model adopts a piecewise linear approximation to estimate the capital investment costs for three different levels of facility capacity. Therefore, annual capital related costs are linearly dependent on the sum of fixed (s_v^l) and variable (s_f^l) capital related costs at every level of capacity at each potential location (Eq. 4.19). The binary decision variable $o_s^{j,l}$ controls the capacity level *l* of the CSP facility located in county j, and the variable $p^{j,l}$ represents the specific capacity of the CSP in county *j* at the capacity level l. The detailed piecewise linear approximation equations were provided in Lin et al. (2013).

$$C_S = C_{S_o} + C_{S_c} \tag{4.17}$$

$$C_{S_o} = s_{op} \times \sum_j \sum_l p^{j,l} \tag{4.18}$$

$$C_{S_c} = \alpha \times (\sum_j \sum_l s_v^l \times p^{j,l} + s_f^l \times o_s^{j,l}) \tag{4.19}$$

Considering the mass balance, the CSP capacity in county *j* should be equal to the total amount of biomass transported to county j from all supply sites in scenario s (Eq. 4.20). The model also considers the biomass loss at the CSP stage that would affect the biomass outlet from CSP to biorefinery ($f^{j,k,s}$) (Eq. 4.21). Biomass loss rate (β) is an input parameter decided by users.

$$\sum_i f^{i,j,s} = \sum_l p^{j,l} \tag{4.20}$$

$$\sum_k f^{j,k,s} \leq \sum_l p^{j,l} \times (1 - \beta) \tag{4.21}$$

Biorefinery

Similar to CSP facilities, the costs related to biorefineries (C_E) are composed of annual biorefinery operating costs (C_{E_o}) and annual biorefinery capital costs (C_{E_c}) (Eq. 4.22). In this study, it is assumed that the unit operating costs for a biorefinery (e_{op}) are constant for any capacity. Therefore, annual operating costs are linearly dependent on the demand of processed biomass for ethanol production at biorefineries (Q), which is an input parameter decided by users (Eq. 4.23). Annual biorefinery capital costs have a linear relationship (α = 13.7% in the current study) with biorefinery capital investment costs, which are the sum of fixed and variable capital related costs at every level of capacity at each potential location (Eq. 4.24). The binary variable $o_e^{k,l}$ indicates whether there exists a biorefinery at capacity level l located in county k, and the variable $q^{k,l}$ represents the biorefinery with the capacity at level l located in county k. Similar to CSP facilities, a piecewise linear approximation method for biorefinery capacity and capacity level identification was implemented. Regarding mass balance, the amount of all the preprocessed biomass flow into the biorefinery located in county k from all CSPs in scenario s should be equal to the biorefinery facility capacity (Eq. 4.25). The total capacity of all biorefineries should meet the given demand of processed biomass for ethanol production (Eq. 4.26).

$$C_E = C_{E_o} + C_{E_c} \tag{4.22}$$

$$C_{E_o} = e_{op} \times Q \tag{4.23}$$

$$C_{E_c} = \alpha \times (\sum_j \sum_l e_v^l \times q^{k,l} + e_f^l \times o_e^{k,l}) \tag{4.24}$$

$$\sum_j f^{j,k,s} = \sum_l q^{k,l} \tag{4.25}$$

$$\sum_{k}\sum_{l} q^{k,l} = Q \tag{4.26}$$

4.3 CASE STUDY

Two spatial and temporal varying coeeficient models GTWR and Bayesian STVC are employed to study the variability of corn yields in response to weather covariates including temperature and precipitation at the county scale in six Corn Belt states in the Midwestern United States. Based on the historical corn yield and weather information, the impacts of temperature and precipitation on the variation of corn yield over space and time are comprehensively investigated. The goal is to understand whether such impacts follow any spatially as well as temporally non-stationary process. After studying the relationship between weather and corn yield, different weather scenarios are simulated to capture the variability of corn yield. Along with two other main uncertain factors on the supply side which include farmer participation rate and corn stover collectable rate, 36 biomass supply scenarios are constructed in the two-stage stochastic programming based optimization.

4.3.1 Data and Variables

Our experiments focus on 606 counties in the Corn Belt states of Illinois, Indiana, Iowa, Kansas, Missouri, and Nebraska. County-average corn yield data for the period of 1981 to 2015 are collected from U.S. Department of Agriculture's National Agricultural Statistics Service (2017). The 35-year average corn yields range from 53.4 to 158.8 bushels per acre for each county (Figure 4.1). When aggregated at the state level, all six states show increasing trends of corn yield over the past 35 years except for a significant hit by severe drought in year 2012 (Figure 4.2). The increasing trends of corn yields can be explained by improved genetics and better management practices. From the available data, about 6.14% of the yield information are missing for multiple

reasons like, no corn is planted for the county (e.g. St. Louis City), or surveys are not conducted in multiple years for some counties. To address this issue, we replace the missing data with the mean yield value of adjacent counties for estimation.

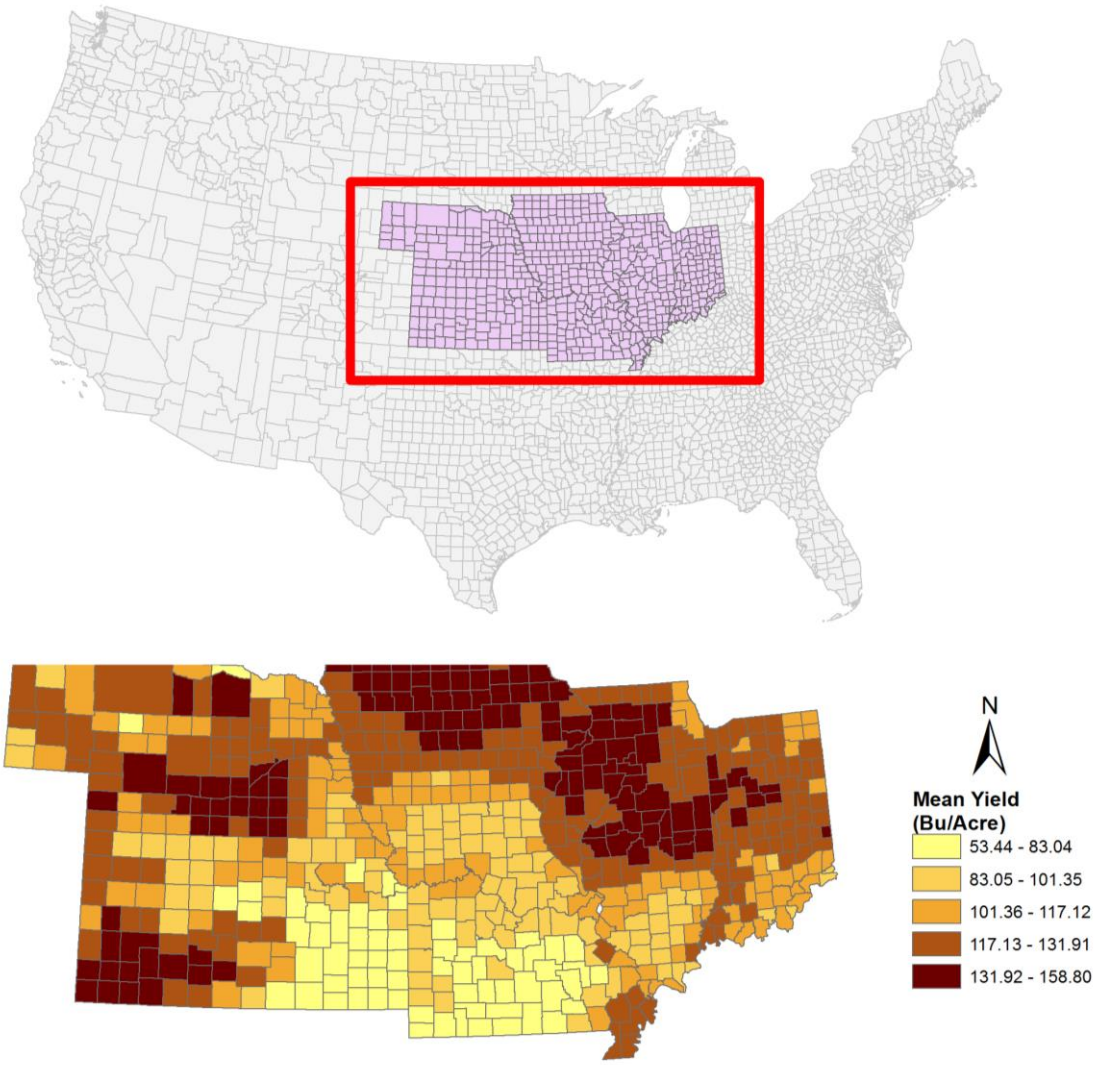


Figure 4.1 Study area and county average corn yield from 1981-2015.

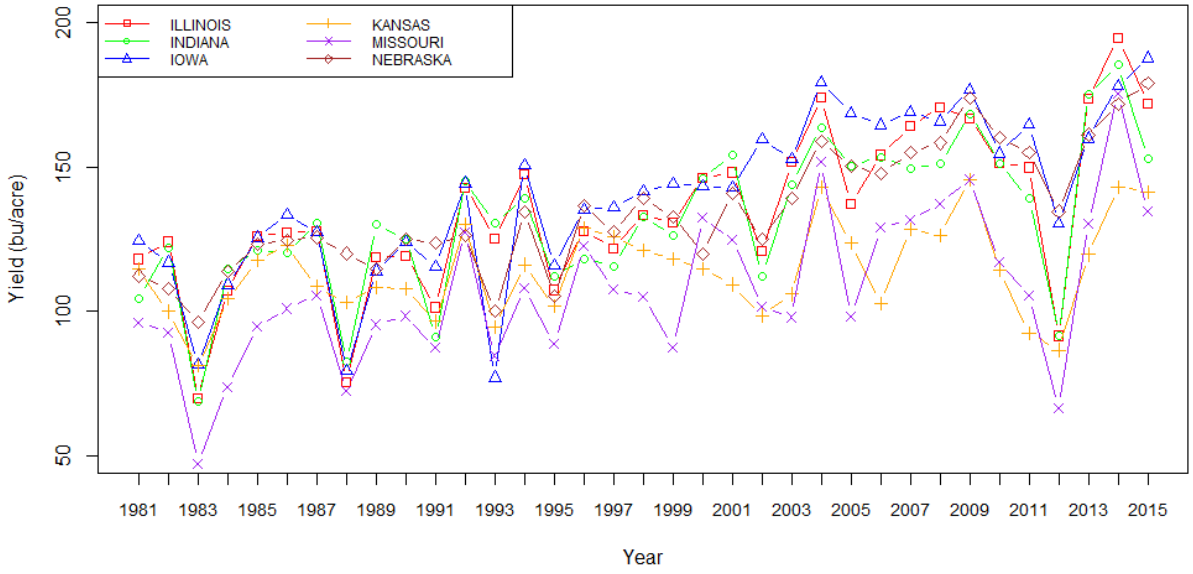


Figure 4.2 Corn yield trends at state level from 1981-2015.

Weather data are derived from PRISM Climate Group (2004). We acquire this data from the Applied Climate Information System API (2017). In particular, maximum daily temperature and minimum daily temperature, and daily precipitation information are used. Instead of using both maximum and minimum temperature information as covariates, we calculate the indicator of growing degree days (GDD), which often serves as a simple single-dimensional summary for describing crops' exposure to heat. GDD is calculated by taking an average of the daily minimum and maximum and subtracting a base temperature value:

$$GDD = \frac{T_{MAX} + T_{MIN}}{2} - T_{BASE} \tag{4.27}$$

There are different methods for calculating GDD. In this study, the most commonly used method in calculating GDD for corn is employed (McMaster and Wilhelm 1997). Constraints on maximum and minimum temperatures are applied for the purpose of eliminating the effect of low or high

temperatures that prevent or retard the growth of corn. More specifically, before entering temperature data into Eq. (4.27), T_{MAX} and T_{MIN} are set equal to T_{BASE} if less than T_{BASE}, and set equal to T_{UT} when greater than T_{UT}. T_{BASE} and T_{UT} are set equal to 10 °C and 50 °C respectively.

Since the PRSIM weather data are served as continous surfaces (spatial raster layers) with 4km x 4km resolution, zonal statistics are used by overlapping the data with county boundaries to estimate the average values of GDD and precipitation at the county scale. GDD and precipitation data are aggregated over the major growing season from 1981 to 2015. The spatial variation of average GDD and precipitation over the observed time period are presented in Figure 4.3 and Figure 4.4. The growing season is considered from the start of May to the end of August, which is the typical growing period for corn.

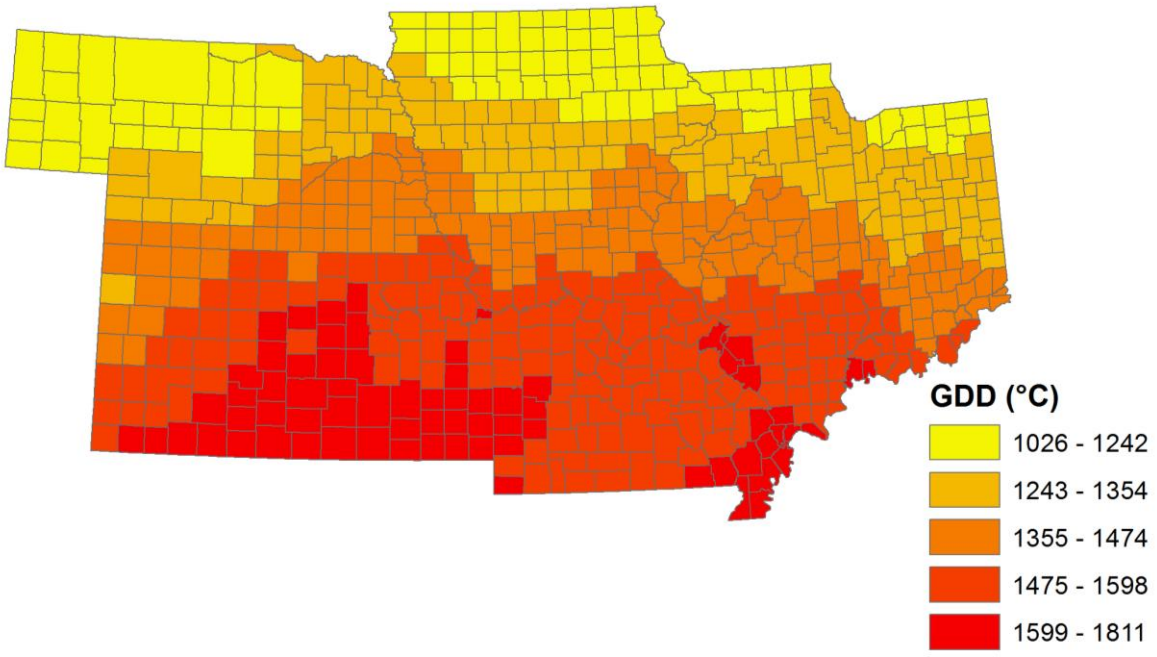


Figure 4.3 Accumulated GDD (°C) from May to August averaged by year from 1981-2015.

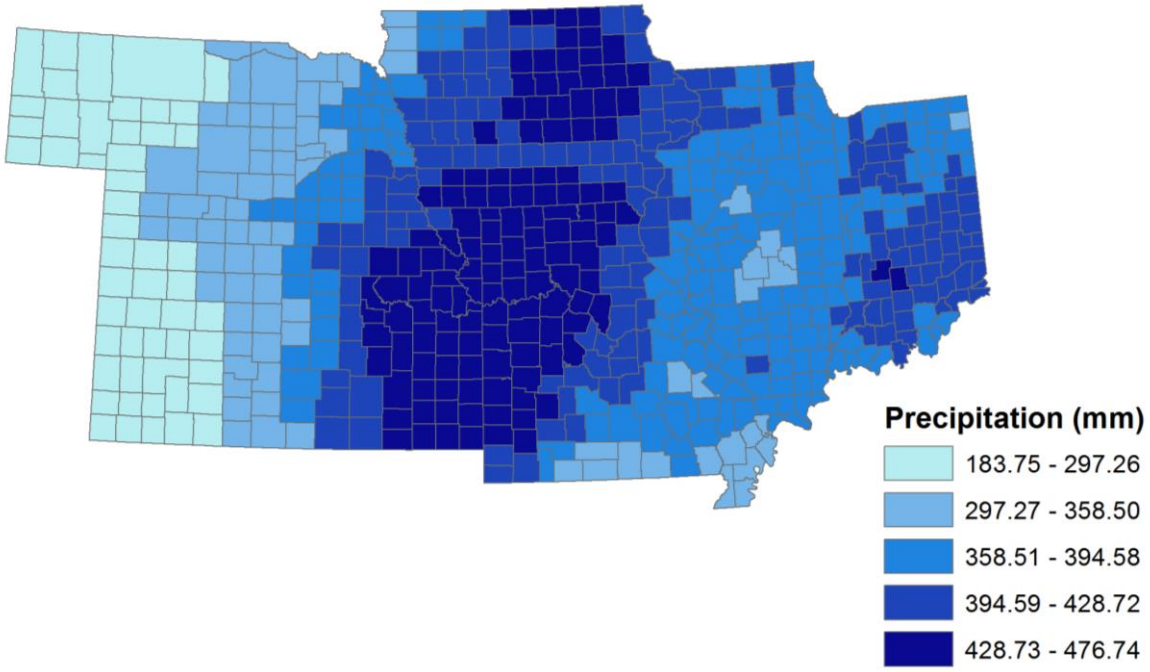


Figure 4.4 Accumulated precipitation (mm) from May to August averaged by year from 1981-2015.

4.3.2 Model Specification and Parameter Estimation

In the specification for Bayesian STVC mode, by combining Equation (4.7)-(4.10), the relationship between corn yields and weather variables including GDD and precipitation (PCPN) are represented as follows:

$$\begin{aligned} Yield_{s,t} = \beta^{Intercept} + \beta^{GDD} GDD_{s,t} + \beta^{PCPN} PCPN_{s,t} + \\ \beta_s^{Intercept} + \beta_{s,t}^{GDD} GDD_{s,t} + \beta_{s,t}^{PCPN} PCPN_{s,t} + \varepsilon_{s,t} \end{aligned} \tag{4.28}$$

where *s* is the index for county, *t* is the index for year, $\beta^{Intercept}$, β^{GDD}, and β^{PCPN} control the overall mean process of coefficients at the global level, among which $\beta_s^{Intercept}$ represents the county-specific intercept to account for time-invariant spatial heterogeneity, and $\varepsilon_{s,t}$ is the white noise. Here we detrend corn yields and weather variables (Figure 4.5) before running the regression

analysis as suggested by Ray *et al.* (2015). This process removes the fixed trend of linear effects caused by government policies and technological improvements, so we focus on the variation of corn yields explained by weather changes. To consider temporal variation in the spatially non-stationary process, $\beta_{s,t}^{GDD}$ and $\beta_{s,t}^{PCPN}$ are defined as spatiotemporally varying coefficients for GDD and precipitation, and they can be further decomposed as:

$$\beta_{s,t}^{GDD} = \beta_s^{GDD} + \alpha_t^{GDD} \tag{4.29}$$

$$\beta_{s,t}^{PCPN} = \beta_s^{PCPN} + \alpha_t^{PCPN} \tag{4.30}$$

where β_s^{GDD} and β_s^{PCPN} represent the time-invariant spatial random effects and α_t^{GDD} and α_t^{PCPN} represent the location-invariant temporal random effects. All county-specific terms, β_s^{GDD}, β_s^{PCPN} and $\beta_s^{Intercept}$, are considered to follow the ICAR prior defined in Equation (4.7). The temporal changes of GDD and precipitation process are modelled as random walk processes as follows:

$$\alpha_t^{GDD} = \alpha_{t-1}^{GDD} + \varepsilon_t^{GDD} \tag{4.31}$$

$$\alpha_t^{PCPN} = \alpha_{t-1}^{PCPN} + \varepsilon_t^{PCPN} \tag{4.32}$$

where ε_t^{GDD} and ε_t^{PCPN} are white noise terms. By choosing the random walk process, we assume that the local coefficient for GDD or precipitation within a specific year is composed of its coefficient from last year plus a random error.

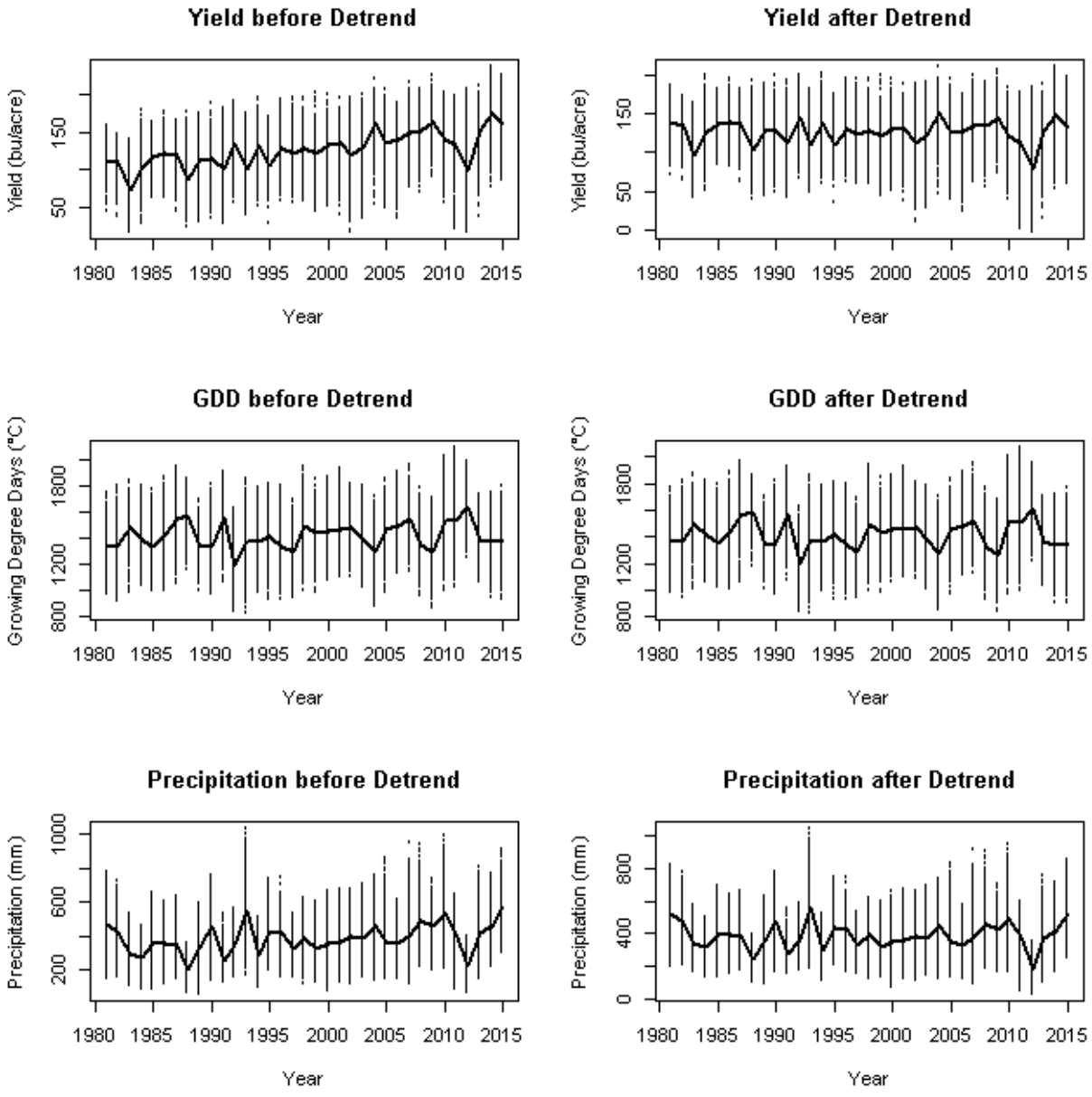


Figure 4.5 Corn yields and weather variables before and after the detrending.

For the purpose of comparison of our analyses, we run the following three alternative models against the data: the OLS model in Equation (4.1), the panel regression model in Equation (4.2), and the GTWR model in Equation (4.5). As suggested by Lobell and Burke (2010), the panel regression model includes a fixed spatial effects term to capture time-invariant heterogeneity, such

as soil quality. In the GTWR model configuration, a spatiotemporal kernel function is applied which consists of mixed spatial and time-decay bandwidths with optimal bandwidth selected by cross validation method as mentioned in Huang et al. (2010).

Following the estimation of SVC, we use a Bayesian method to estimate our STVC model. The likelihood of the crop yields in Equation (4.28) is expressed as:

$$f(\mathbf{Yield}|\Theta) = \prod_s \prod_t N(Yield_{s,t} | GDD_{s,t}, PCPN_{s,t}, \beta^{Intercept}, \beta^{GDD}, \beta^{PCPN}, \beta_s^{Intercept}, \beta_s^{GDD}, \beta_s^{PCPN}, \alpha_t^{GDD}, \alpha_t^{PCPN}, \sigma^2) \quad (4.33)$$

where Θ is a set of all the hyper-parameters included in the model. The prior specification for our model is as follows. $\beta_s^{Intercept}, \beta_s^{GDD}$ and β_s^{PCPN} have ICAR prior. α_t^{GDD} and α_t^{PCPN} follow the random walk process. A vague normal $N(0,10^6)$ is used for $\beta^{Intercept}$, β^{GDD}, β^{PCPN}. A vague inverse gamma, $IG(1,0.01)$, is set as the prior distribution for all white noise parameters.

Given the complexity of posterior distributions in hierarchical models, both Gibbs sampler and Metropolis Hasting algorithms are used in Markov chain Monte Carlo (MCMC) to generate posterior samples (Gilks *et al.* 1995). MCMC sampling methods provide a general approach to fiting complex hierarchical models in a Bayesian framework (Cressie and Wikle 2015). MCMC generates values of model parameters from a set of Markov chains after a number of steps until they converge. The converged values are then used as a sample from the posterior distirtion to enable Monte Carlo estimation of the joint, marginal, and conditional posterior distributions desired for inference. In our study, two chains are initialized and posterior distributions for each model parameters were estimated after 10,000 iterations with the first 2,000 iterations discarded as the burn-in period.

While parameters of the STVC model are estimated, the posterior mean of corn yield can also be derived to quantitatively represent the predicted corn yield when new weather data are available.

Through this approach, the uncertainty of corn yield can be represented by simulating different weather scenarios.

4.3.3 Scenarios Construction in Stochastic Programming

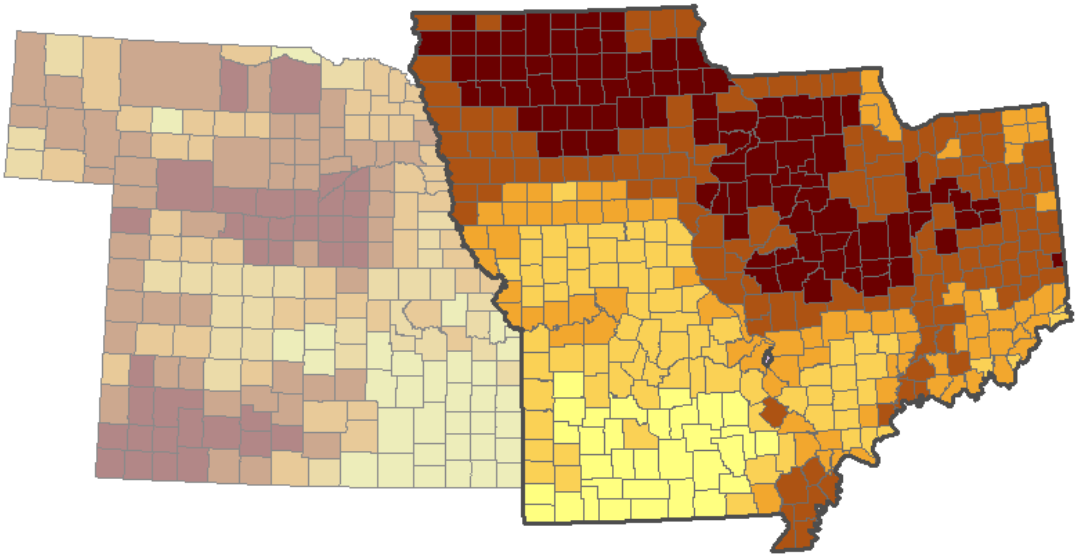


Figure 4.6 Candidate corn stover supply region (four states in Illinois, Indiana, Iowa and Missouri).

In the stochastic optimization of the biomass (corn stover) supply chain, four states in Illinois, Indiana, Iowa and Missouri are considered as candidate biomass supply counties (Figure 4.6). Three major sources of uncertainty which impact on the corn stover availability are considered in this study – corn yield, corn stover collectable rate, and farmer participation rate. Corn yield is mainly affected by weather dynamics during the growing season. Nine weather scenarios are simulated through varying growing degree days (GDD) and precipitation by one standard deviation (±12.9 for GDD, ±32.5 for precipitation) from their distribution in historical data over the past 35 years (Figure 4.7). The posterior means of corn yield are generated based on the STVC model.

The corn stover collectable rate represents the amount of residuals (i.e. stalks, leaves, and cobs) that are collected after corn harvesting, and it is related to residue management practice. Two collectable corn stover rates are selected to represent lower residue management (0.3) and higher residue management (0.5). The farmer participation rate indicates the percentage of farmers who are willing to sell the collected corn stover for biofuel production purpose. Two farmer participation rates are selected to represent lower farmer interest (24%) and higher farmer interest (36%). The probability distribution of farmer participate rate and residual management are based on the survey results from Iowa (Tyndall et al., 2011). In total, nine weather scenarios, two levels of residue management and three levels of farmer interest levels are incorporated to present the 36 scenarios of biomass supply scenarios (Table 4.1).

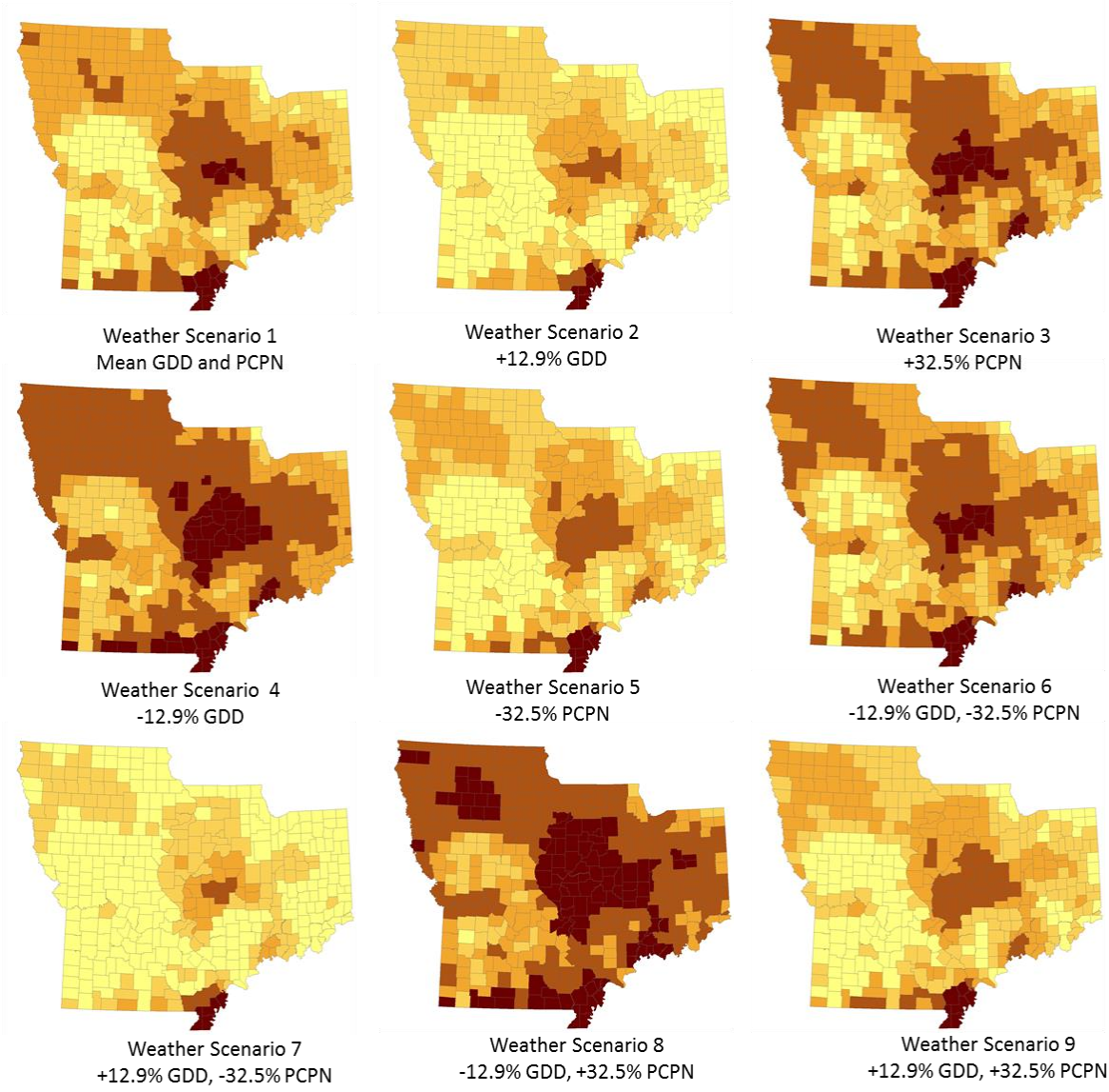


Figure 4.7 Corn yield under nine weather scenarios with varying growing degree days (GDD) and precipitation (PCPN) at one standard deviation

Table 4.1 Biomass provision scenarios with its probabilities

Scenarios	Weather scenarios	Farmer participate rate	Collectable corn stover rate	Probability
S1	No Change (30%)	0.24 (80%)	0.3 (70%)	16.8%
S2			0.5 (30%)	7.2%
S3		0.36 (20%)	0.3 (70%)	4.2%
S4			0.5 (30%)	1.8%
S5	+12.9% GDD (10%)	0.24 (80%)	0.3 (70%)	5.6%
S6			0.5 (30%)	2.4%
S7		0.36 (20%)	0.3 (70%)	1.4%
S8			0.5 (30%)	0.6%
S9	+32.5% PCPN (10%)	0.24 (80%)	0.3 (70%)	5.6%
S10			0.5 (30%)	2.4%
S11		0.36 (20%)	0.3 (70%)	1.4%
S12			0.5 (30%)	0.6%
S13	-12.9% GDD (10%)	0.24 (80%)	0.3 (70%)	5.6%
S14			0.5 (30%)	2.4%
S15		0.36 (20%)	0.3 (70%)	1.4%
S16			0.5 (30%)	0.6%
S17	-32.5% PCPN (10%)	0.24 (80%)	0.3 (70%)	5.6%
S18			0.5 (30%)	2.4%
S19		0.36 (20%)	0.3 (70%)	1.4%
S20			0.5 (30%)	0.6%
S21	-12.9% GDD -32.5% PCPN (5%)	0.24 (80%)	0.3 (70%)	2.8%
S22			0.5 (30%)	1.2%
S23		0.36 (20%)	0.3 (70%)	0.7%
S24			0.5 (30%)	0.3%
S25	+12.9% GDD -32.5% PCPN (10%)	0.24 (80%)	0.3 (70%)	5.6%
S26			0.5 (30%)	2.4%
S27		0.36 (20%)	0.3 (70%)	1.4%
S28			0.5 (30%)	0.6%
S29	-12.9% GDD +32.5% PCPN (10%)	0.24 (80%)	0.3 (70%)	5.6%
S30			0.5 (30%)	2.4%
S31		0.36 (20%)	0.3 (70%)	1.4%
S32			0.5 (30%)	0.6%
S33	+12.9% GDD +32.5% PCPN (5%)	0.24 (80%)	0.3 (70%)	2.8%
S34			0.5 (30%)	1.2%
S35		0.36 (20%)	0.3 (70%)	0.7%
S36			0.5 (30%)	0.3%

The biofuel demand for all scenarios considered was constant at 378.5 million Liters per year, or 100 million gallons per year. Annual demand for corn stover at biorefinery gate was estimated at 1.25 million Mg, assuming an ethanol conversion rate from corn stover at 302.5 Liters per Mg.

4.4 RESULTS AND DISCUSSIONS

4.4.1 Relationships between Weather and Crop Yields

Parameters estimated from all the models considered in this study are summarized in Table 4.2. The slope coefficient for GDD and precipitation can be interpreted as changes of corn yield (bushel/acre) per unit of accumulated GDD change (°C), or per unit of accumulated precipitation change (mm) over the growing season for a specific county. The results from the OLS model show that corn yields are negatively associated with temperature but have a slightly positive relationship to precipitation although the coefficient for precipitation is not significant at 0.05 level. Although residuals of OLS model are normally distributed, strong positive spatial autocorrelation is detected by Moran's *I* index with an average value around 0.471 and all *p*-values less than 0.05 over 35 years (Table 4.2).

Compared to OLS, the spatial panel regression model shows a more significant relationship between precipitation and crop yields with a positive coefficient around 0.017. With the expense of additional parameters of spatially varying intercepts in its mean structure, the spatial panel model accounts for 54% of crop yield variation, compared to only 20.7% for OLS. Additionally, residuals in the spatial panel regression model are less correlated than those in OLS, giving a significant decrease of Moran's *I* (Table 4.2).

Table 4.2 Summary of parameters for OLS, Panel Regression with Spatial Fixed Effects, GTWR, and STVC Models on Corn Yields and Weather data for 606 Midwestern US Counties, 1981 – 2015.

Parameters	OLS	Panel Regression			GTWR			STVC		
		Min	Mean	Max	Min	Mean	Max	Min	Mean	Max
Intercept	232.8 (1.698)	187.0	241.5	314.1	103.7	231.5	281.3	110.6	234.7	266.5
Slope for GDD	-0.075 (0.001)		-0.086 (0.001)		-0.088	-0.051	0.049	-0.078	-0.042	0.031
Slope for precipitation	0.001^{NS} (0.001)		0.017 (0.001)		-0.155	0.017	0.193	-0.184	0.012	0.188
Adjusted R-squared	0.207		0.540			0.715			0.749	
Cross validation RMSEs	29.84		20.66			21.35			21.54	
Average Moran's *I*	0.471		0.168			0.126			0.113	

Notes: One digit after decimal point is displayed for intercept, three digits after decimal point are displayed for coefficients and adjusted R-squared, and two digits after decimal points are displayed for RMSEs. Standard errors are shown in parentheses. Coefficients denoted with NS represent they are not significant at the 0.05 confidence level. The significance of coefficients is represented in Figure 4.8 All the parameters for the STVC model are posterior means. Adjusted R-squared for STVC is also calculated based on posterior mean of fitted values.

Non-stationary coefficients are computed by the two competitive models GTWR and STVC respectively, and the GTWR estimates are generally more variable as shown in Table 4.2, except for the lower bound slope for precipitation coefficient. For modeling fitting performance, GTWR and STVC models further improve the adjusted R-squared value to 71.5% and 74.9% with the consideration of spatiotemporally varying association between crop yields and weather factors respectively. Mean values of GDD and precipitation coefficients follow the same trend as those estimated by OLS and spatial panel regression models, although slopes in STVC are smoother compared to those in GTWR. Since both GTWR and STVC achieve quite similar coefficient estimations, but STVC has more flexibility on configuring spatial and temporal priors, the following discussion is more focused on the STVC results and the coefficients are applied to combined with the subsequent stochastic programming model.

The spatial non-stationarity of coefficients estimated by the STVC model is first examined by averaging out their change over time. The impact of GDD and precipitation are complementary to each other in the fact that regions with higher GDD coefficient values are more likely to have lower precipitation effects, and vice versa (Figure 4.8). Overall, we observe that temperature, reflected by GDD, tends to have negative effects on corn yields in warmer regions and positive effects in cooler regions. By comparing the precipitation coefficient surface with the map of irrigated corn harvested acres percentage (Figure 4.9), it is found that some heavily irrigated land tends to have insignificant or even negatively significant precipitation effects, such as in Nebraska and Kansas. This finding agrees with what has been previously discovered in GWR methods (Sharma *et al.* 2011; Cai et. al 2014) that crop yields could be less related to precipitation in regions with better irrigation systems.

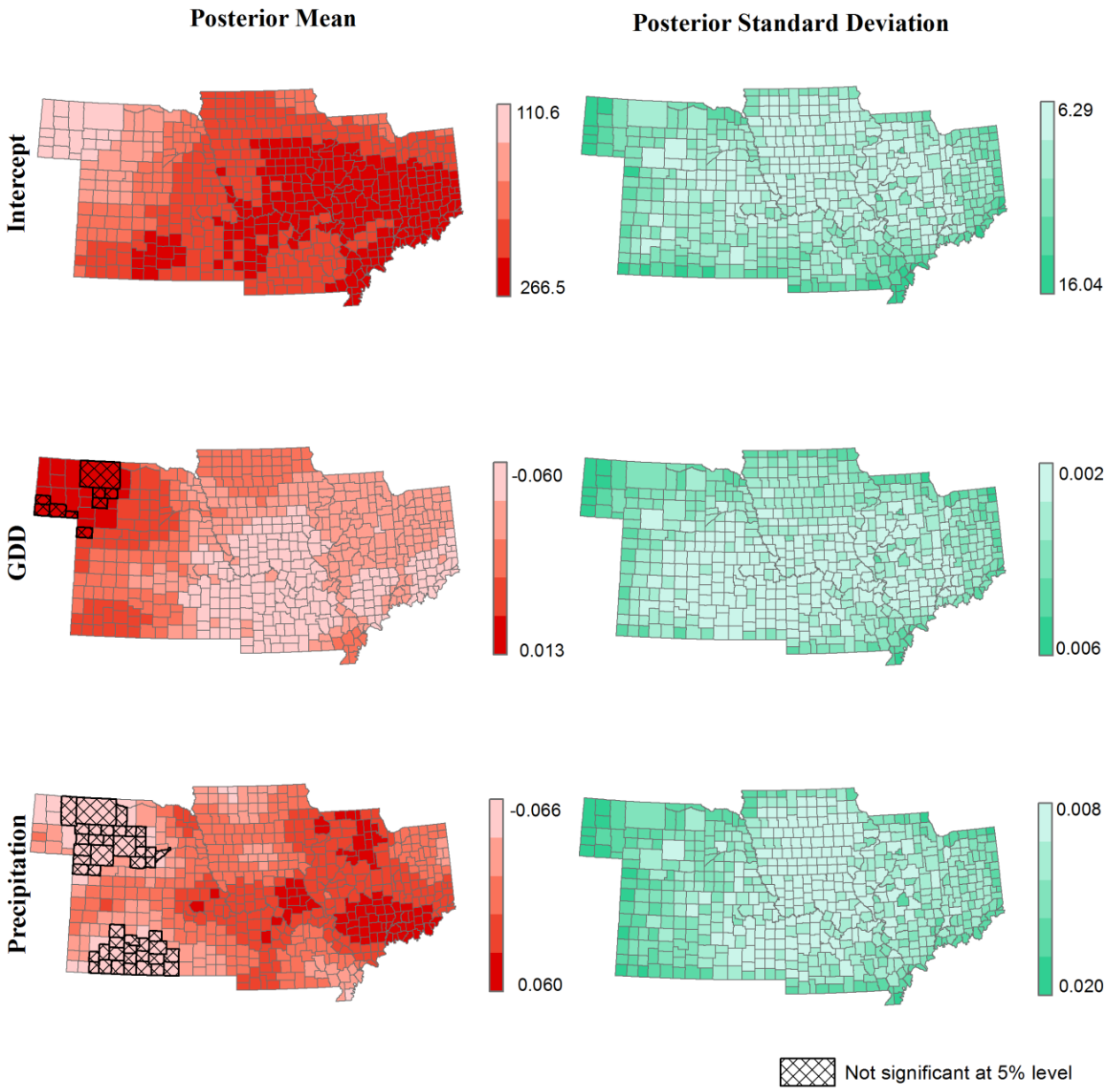


Figure 4.8 STVC estimated coefficients for the intercept, GDD, and precipitation averaged from 1981-2015.

The standard deviation of the posterior distribution of coefficients in STVC is provided in Figure 4.8 as well to visually examine the uncertainty. Edge effects are obviously detected because of the sudden cut off at boundaries. To show the significance of local coefficients, we mark those

counties in which coefficients are not significantly different from zero on the map (Figure 4.8 We use the range of 2.5%-97.5% quantile of the posterior distribution to check if zero is included. If yes, those counties are considered as not significantly different from zero. Temporally varying coefficients are also included in testing the significance of parameters in STVC, which generates 35 parameter surfaces to represent the significance of parameters in different years. For the convenience of visualization, we mark counties with a crosshatch pattern if their coefficients are not significant for more than 21 years. Overall, significant spatially varying coefficients are captured by STVC.

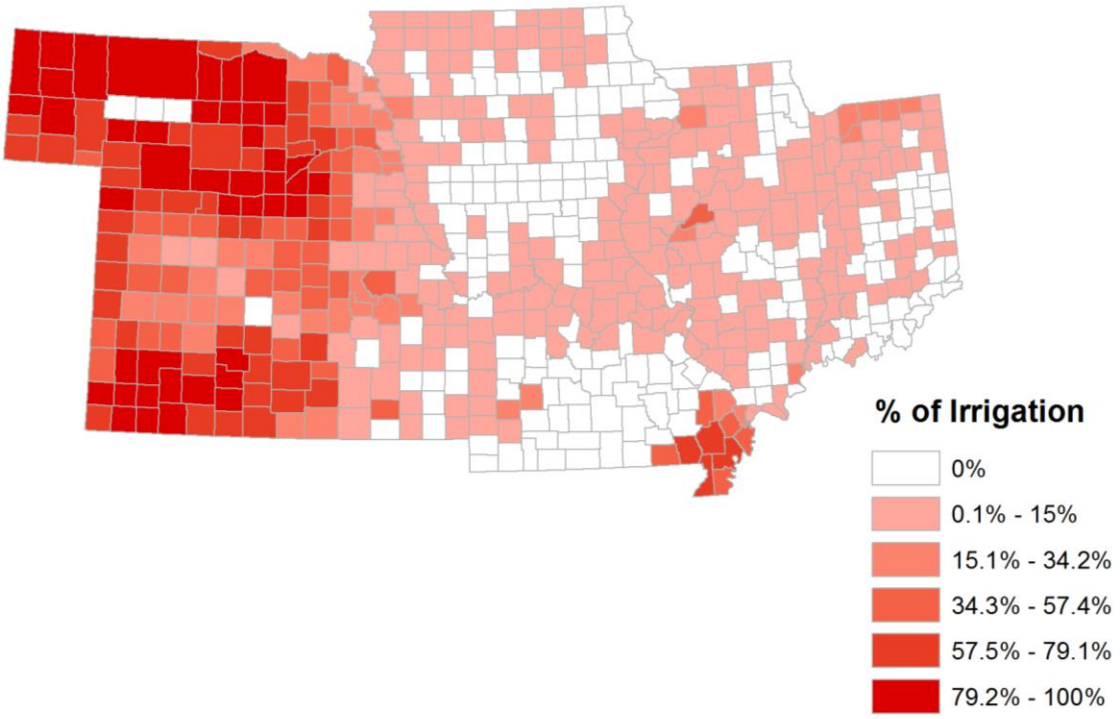


Figure 4.9 Percentage of harvested acres for irrigated corn based on averaged data for the year 1997, 2002, 2007, and 2012 from USDA National Agricultural Statistics Service.

The STVC model accounts for the temporal variation of coefficients by introducing a random walk process to allow spatially varying coefficients to change over time. As shown in Figure 4.10,

changes in GDD and precipitation coefficients are estimated based on random walk priors. Similar to what has been presented spatially, the change of coefficients for GDD and precipitation are negatively correlated. We further investigate into the potential reasons for the temporal variation of coefficients. For example, in the year 1988 the coefficient of precipitation increases by nearly 0.12. Given that the mean value is only 0.012 (Table 4.), this increase is dramatic. By comparing to the average GDD and precipitation (Figure 4.11), it is found that the average precipitation in 1988 is the lowest among the years we included. That means the marginal increase of corn yield is significant in conditions with insufficient rains. With the same motivation, we examined the year with the lowest GDD coefficient, which is 2012. Correspondingly, the accumulated GDD in that year is the highest, which means high temperatures will accelerate the rate of decreasing in corn yields. There is one thing that could not be explained well based on the current results. In the year 2011, the model estimated a high precipitation coefficient and a low GDD coefficient, however, the accumulated precipitation and GDD, especially the precipitation value, are not obviously deviated from their mean values. We suspect that this variation might be caused by non-weather related factors. The ability to estimate temporally varying coefficients is a significant contribution of the model we developed, and this could further lead to the study of the non-linear response of crop yields to weather variations.

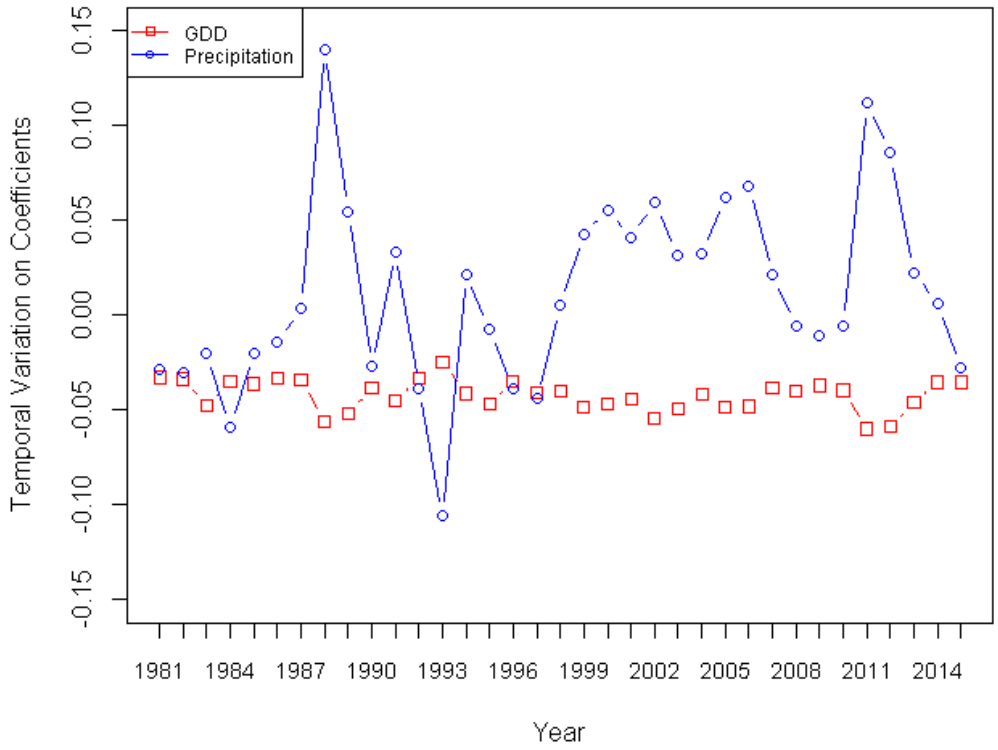


Figure 4.10 Temporal variations on coefficients for GDD and precipitation.

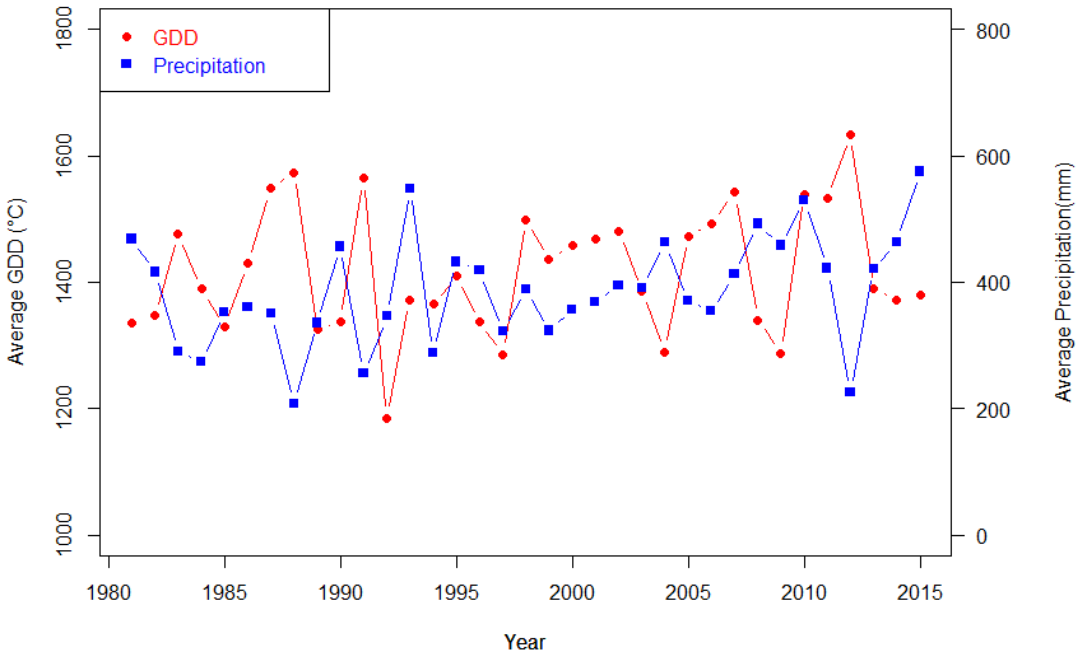


Figure 4.11 Average GDD and precipitation in the study area from 1981 to 2015.

In addition to using the adjusted-R values to compare model fitting performance, we plot the fitted corn yield values against their true values for the following six counties, namely, Coles in Illinois, Franklin in Kansas, Brown in Kansas, Knox in Nebraska, Rooks in Kansas, and Barry in Missouri (Figure 4.12). These counties are selected because they have the worst fitting performances in terms of deviation from true values reported by the OLS model. According to the figure, STVC performs better than the other models in general. It captures some low yield values that the other models failed to fit well (e.g., Franklin-Kansas in 2012, Rooks-Kansas in 1993), while high deviations (e.g., Brown-Kansas in 2012) were also observed sometimes.

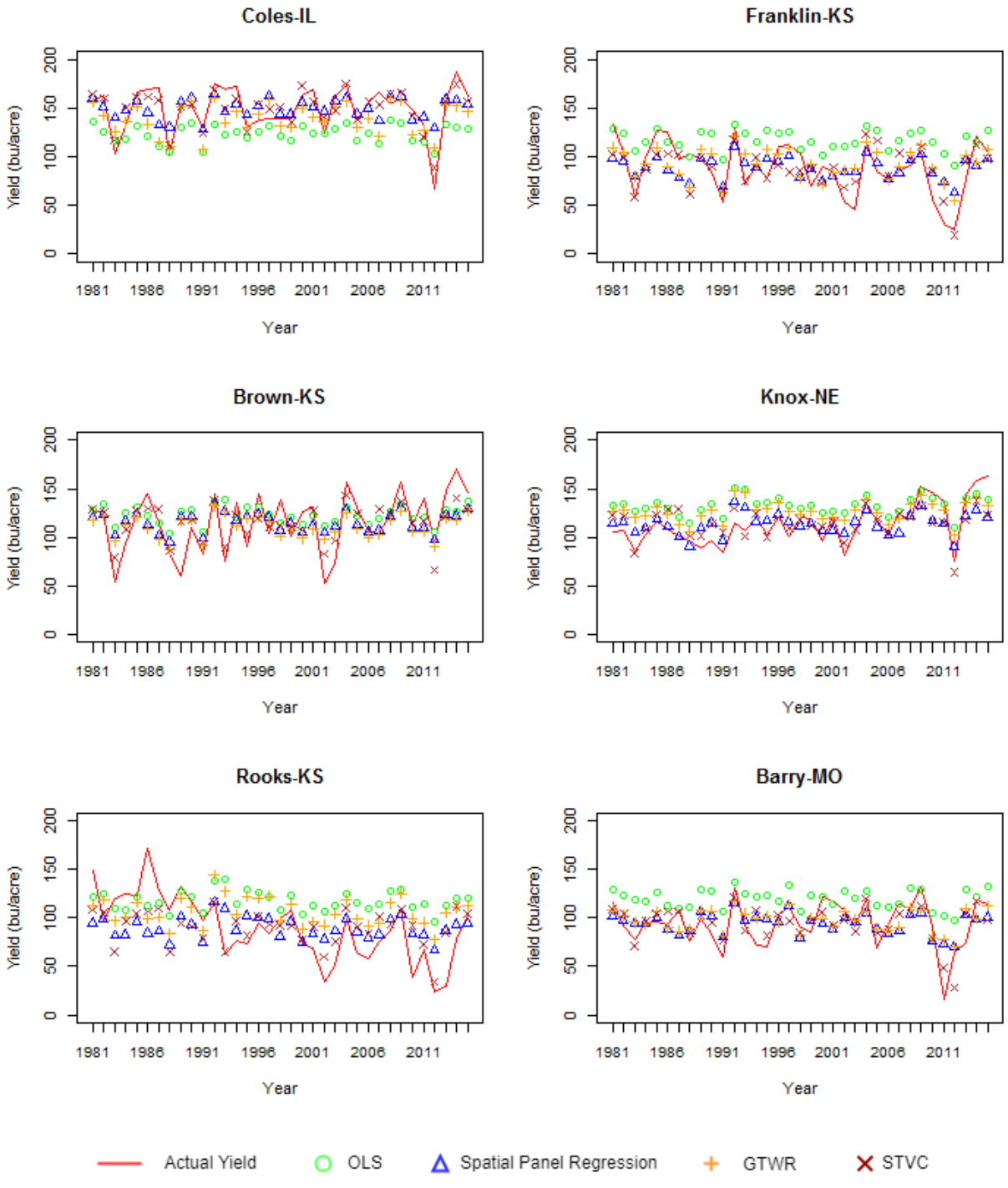


Figure 4.12 Comparison of model fitting performance in selected counties.

4.4.2 Stochastic Optimization Results

Considering all possible scenarios together, the stochastic optimization model results showed that the expected minimal biomass-ethanol production costs are $0.763 per Liter under the optimal supply chain configuration (Table 4.3). Biorefinery related costs account for almost half of the cost with $0.372 per Liter, and followed by biomass provision costs, transportation, and CSP related costs. CSP costs and biorefinery related costs are constant for all scenarios because they are only subject to change in second-stage decisions (facility location and numbers) and the demand of bioethanol. Considering the dynamics of 36 scenarios, the stochastic optimization model chose to build two CSP facilities located in DeWitt County and Logan County in Illinois and one biorefinery facility located in DeWitt County Illinois, as the optimal supply chain configuration (Figure 4.13). The optimal facility locations are first-stage decisions in stochastic modeling, which is constant for all scenarios. Given the uncertainty in biomass availability, up to 45 supply counties can be selected in 36 scenarios as second-stage decisions.

The minimal unit biomass-ethanol production costs of each scenario change from $0.727 to $0.812 per Liter. The biomass provision costs and transportation costs are the major contributions, which vary from $0.220 to $0.239 per Liter and from $0.088 to $0.154 per Liter, respectively. These cost changes are the results of changing biomass availability in each scenario. Spatially, the number of selected supply counties and transportation patterns changes by different scenarios, as a result of changing annual biomass availability. For example, Figure 4.14 compares two extreme scenarios with least and most corn stover availability respectively. Scenario 25 has the lowest corn stover availability with worst weather scenario for corn yield, lower farmer participation rate and lower collectable corn stover rate, while Scenario 32 has the highest corn stover availability with best weather scenario for high corn yield, higher farmer participation rate and higher collectable corn stover rate. Higher biomass availability reduces the number of selected biomass supply

counties. Accordingly, the number of required supply counties in Scenario 32 (10) is almost one-fifth of the number required by Scenario 25 (45).

Table 4.3 Stochastic optimization results for 36 scenarios of biomass supply. S25 has the highest unit bioethanol production cost while S32 has the unit bioethanol production cost All the numbers related to cost are in terms of $ per Liter.

Scenarios	Unit production cost	Provision cost	CSP cost	Biorefinery cost	Transportation cost
1	**0.797**	0.239	0.047	0.372	0.139
2	**0.748**	0.22	0.047	0.372	0.109
3	**0.771**	0.238	0.047	0.372	0.114
4	**0.733**	0.22	0.047	0.372	0.094
5	**0.808**	0.239	0.047	0.372	0.15
6	**0.757**	0.221	0.047	0.372	0.117
7	**0.78**	0.238	0.047	0.372	0.123
8	**0.739**	0.22	0.047	0.372	0.1
9	**0.789**	0.238	0.047	0.372	0.132
10	**0.744**	0.22	0.047	0.372	0.105
11	**0.765**	0.237	0.047	0.372	0.109
12	**0.73**	0.22	0.047	0.372	0.091
13	**0.786**	0.238	0.047	0.372	0.129
14	**0.743**	0.22	0.047	0.372	0.104
15	**0.763**	0.237	0.047	0.372	0.107
16	**0.729**	0.22	0.047	0.372	0.09
17	**0.8**	0.239	0.047	0.372	0.142
18	**0.75**	0.22	0.047	0.372	0.111
19	**0.774**	0.238	0.047	0.372	0.117
20	**0.735**	0.22	0.047	0.372	0.096
21	**0.804**	0.239	0.047	0.372	0.146
22	**0.753**	0.22	0.047	0.372	0.114
23	**0.777**	0.238	0.047	0.372	0.12
24	**0.737**	0.22	0.047	0.372	0.098
25	**0.812**	0.239	0.047	0.372	0.154
26	**0.76**	0.221	0.047	0.372	0.12
27	**0.783**	0.238	0.047	0.372	0.126
28	**0.741**	0.22	0.047	0.372	0.102
29	**0.783**	0.238	0.047	0.372	0.126
30	**0.741**	0.22	0.047	0.372	0.102
31	**0.762**	0.237	0.047	0.372	0.106
32	**0.727**	0.22	0.047	0.372	0.088
33	**0.792**	0.238	0.047	0.372	0.135
34	**0.746**	0.22	0.047	0.372	0.107
35	**0.767**	0.237	0.047	0.372	0.111
36	**0.731**	0.22	0.047	0.372	0.092
Overall	**0.763**	0.230	0.047	0.372	0.115

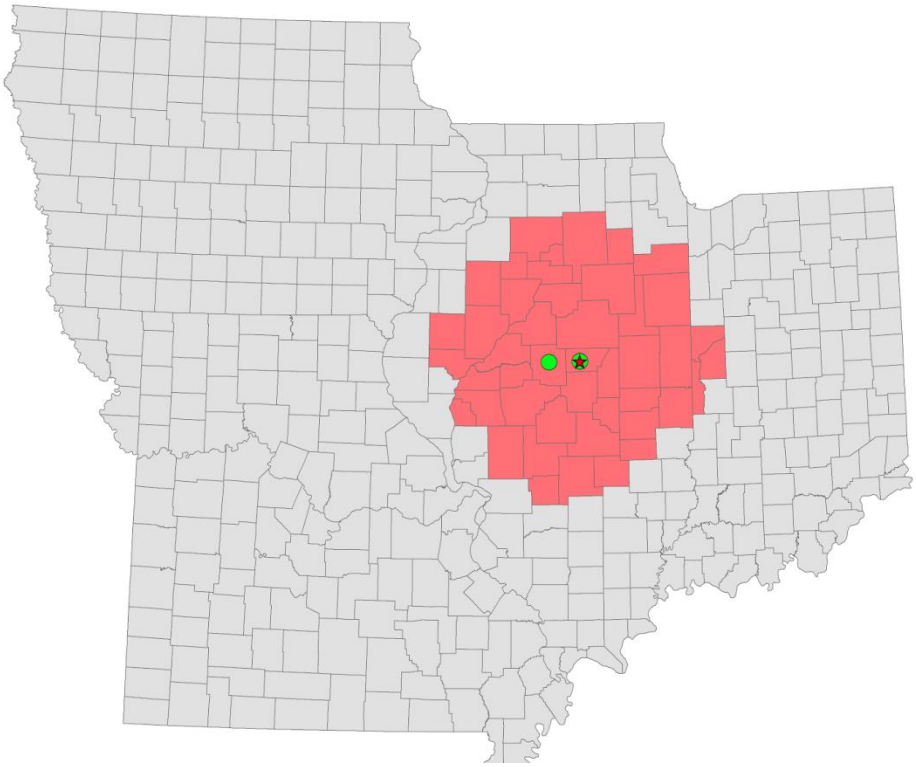
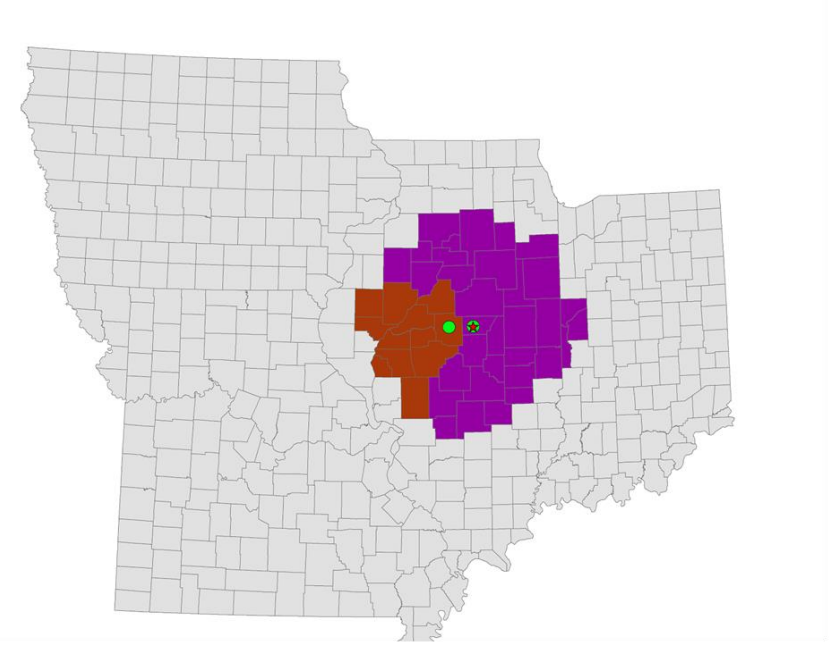
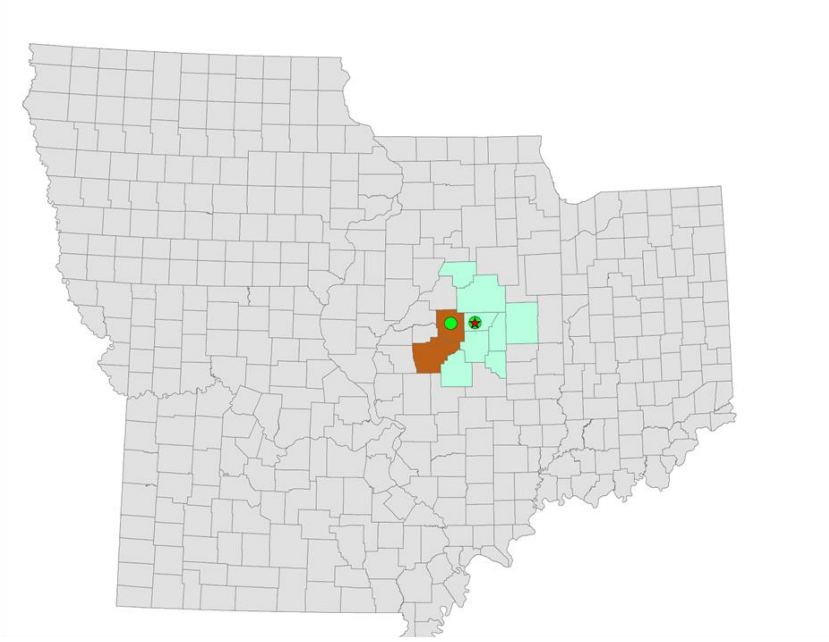


Figure 4.13 Results of stochastic programming. Red star represents the selected location for biorefinery (DeWitt, Illinois); Green circles represent the selected locations for CSP (DeWiit, Illinois and Logan, Illinois); Pink regions represent selected supply counties in all 36 scenarios



S25: +12.9% GDD, -32.5% PCPN, farmer participation rate (0.24), collectable corn stover rate (0.3)



S32: -12.9% GDD, +32.5% PCPN, farmer participation rate (0.36), collectable corn stover rate (0.5)

Figure 4.14 Results of stochastic programming scenarios for highest unit bioethanol production cost (Scenario 25) and lowest unit bioethanol production cost (Scenario 32)

4.5 CONCLUSIONS

In this study, we developed a novel Bayesian hierarchical model, namely spatiotemporally varying coefficient (STVC) model, to investigate the non-stationary relationship between weather and corn yields in the Midwestern US counties for the period of 1981 to 2015. The model treats regression effects as spatially and temporally correlated processes within a Bayesian framework to enable statistical inference on the regression associations. The results showed spatial heterogeneous effects of Growing Degree Days (GDD) and precipitation on crop yields, and meanwhile revealed the change of effects over time.

Compared to alternative models including OLS, spatial panel regression, the STVC model significantly improves the variation of corn yields that can be explained by weather change with the consideration of spatiotemporal non-stationarity. The STVC model achieves similar estimation as that from GTWR but with the identification of the spatial scale of each relationship through its specification of geostatistical priors. More specifically, changes in GDD and precipitation account for 74.9% of the variation in corn yields, with negatively correlated effects over space and time. The marginal increase of corn yields is significant in conditions with insufficient rains, while high temperatures accelerate the decreasing rate of corn yields. Based on the variability of corn yield captured by the STVC model, a two-stage stochastic programming model is developed with the consideration of biomass supply uncertainty to optimize biomass supply chain infrastructure configurations, biomass feedstock (corn stover) provision, and logistics operations

From a methodological standpoint, it is important that the STVC model accommodates non-stationarity of regression coefficients in both spatial and temporal dimensions. The model by design directly captures non-stationary processes in coefficients instead of allowing them to be reflected through the error terms, thus mitigates the issue of spatial autocorrelation in residuals.

The innovation of integrating with stochastic optimization model is to leverage the uncertainty information captured by the STVC model to construct more realistic scenarios in stochastic programming.

From an application perspective, the contribution of our study helps better understand the relationship between corn yields and weather factors. While previous studies have been conducted to examine the spatial heterogeneity of the relationship, this study incorporates the temporal variation of the relationship. By revealing this variation, further exploitation of the data can be carried out with more confirmatory analyses. For example, given the fact that the influence of weather on crop yields is not constant over space and time, further analysis can be conducted to examine the potential factors, e.g. weather and non-weather related, that contribute to the variation. The stochastic optimization of biomass supply chain is demonstrated as an example of how the relationship between corn yields and weather factors can be further leveraged to inform decisions.

One limitation of this work on the statistical modeling part is that we did not fully explore other possible temporal structures or more complicated spatiotemporal nested structures. This is because our focus is placed on incorporating temporal variation to spatially varying coefficient models instead of comparing different spatiotemporal models. However, the latter is still of great interest as future work. One possible extension is to model the temporal variation of weather coefficients as a spatial heterogeneous process. Also, it is interesting to compare different temporal processes, e.g. autoregressive process versus random walk, to better understand temporal structures.

There are opportunities to enhance the current stochastic programming model to represent more real-world scenarios. For example, our ongoing work expands the current model into a multi-period stochastic supply chain optimization model. In this model, the biomass supply chain is planned for a multi-year period (e.g., 10 years) with each year having a fixed increase of bioethanol

production demand. It is also interesting to address the sources of uncertainty from bioethanol production demand perspective. However, when different sources of uncertainty are combined to represent biomass supply and bioethanol demand scenarios, the size of the stochastic optimization problem will be extremely large and thus novel computational approaches to generating optimal solutions need to be developed. Such approaches will likely contribute to the advancement of cyberGIS and HPC-based spatial optimization.

CHAPTER 5
CONCLUDING DISCUSSION

In this thesis, three interrelated studies are presented from complementary perspectives to advance cyberGIS-enabled spatial decision support system informed by uncertainty quantification with applications to biomass-to-biofuel supply chain optimizations. The study (Chapter 2) proposes a cyberGIS-enabled spatial decision support system with a focus on the system architecture and interoperability of data and services tuned for supporting decisions in biomass supply chain optimizations. The system, CyberGIS-BioScope, follows a geodesign approach to spatial decision-making, user environment design, and a service integration approach enabled by the GISolve middleware services. It provides a highly interactive user environment for users and supports the optimization of biomass supply chain and the evaluation of uncertainty through a scalable computational workflow implementation. Benefiting from the transparent access to data, models, and advanced cyberinfrastructure, decision makers can speed up their scientific discovery and knowledge-sharing processes in related collaborative biomass-to-biofuel supply chain research.

Based on the cyberGIS-enabled spatial decision support system for biomass supply chain optimizations proposed in the first study, the second study (Chapter 3) further extends the system with analytical capabilities for performing uncertainty and sensitivity analysis. In the biomass supply chain design and operational planning, model-based approaches are sometimes unreliable when most of the data are fraught with uncertainties. Uncertainty and sensitivity analysis techniques offer an accessible treatment via the quantification of uncertainty propagations and sensitivity measurement. The system described in Chapter 3 serves as a new integrated approach to data management, mathematical modeling, uncertainty and sensitivity analysis, what-if scenario

analysis, and result representation and visualization for the biomass supply chain optimization. Leveraging high-performance computing capabilities provisioned by advanced cyberinfrastructure, this cyberGIS approach enables Monte Carlo based uncertainty and sensitivity analysis for optimization modeling by resolving significant computational intensity. Furthermore, the approach supports dynamic management decisions through what-if scenario analysis responding to uncertain situations in supply chain operations.

A major contribution of the paper described in Chapter 3 is that it bridges the gaps between research, development, and implementation of biomass supply chain optimization under uncertainties. This approach generalizes and streamlines data management, computation, and visualization components, so it is expected to work on biomass supply chain optimization applications customized by different model and data inputs. From the case study of optimizing Miscanthus supply chain in Illinois, United States, the system is effective to capture a range of optimal bioethanol production cost with different sources of uncertainty considered and meanwhile identify the rank of most influential factors to optimal bioethanol production cost, which are valuable information for investment strategies in bioenergy production industry.

The third study (Chapter 4) described in this thesis focuses on a novel statistical modeling approach for uncertainty quantification on the biomass supply side. In the meantime, a two-stage stochastic programming model is developed based on the results of uncertainty analysis to support biomass supply chain optimizations under supply uncertainty in the application of biomass (corn stover) supply chain optimization in six Corn Belts states in the US. The proposed statistical model, namely spatiotemporally varying coefficient (STVC) model, correlates the corn yield with weather information and significantly improves the variation of corn yields that can be explained by weather change with the consideration of spatiotemporal non-stationarity. Compared to a similar

method named geographical and temporal weighted regression (GTWR) in literature, STVC provides more flexibility on defining the structure of spatial and temporal random effects and results in less spatially autocorrelated residuals. It is important that the STVC model accommodates non-stationarity of regression coefficients in both spatial and temporal dimensions to better understand the relationship between corn yields and weather factors. This understanding provides a better way to represent spatiotemporal uncertainty of biomass supply in the context of biomass supply chain optimizations. As opposed to the more traditional approach which applies a global probability distribution to the quantity of biomass supply to generate discrete scenarios in stochastic programming models, this study exploits weather information to generate different scenarios of biomass supply with the assumption that weather information can be represented with certain probability distributions. The stochastic optimization of biomass supply chains is demonstrated as an example of how the relationship between corn yields and weather factors can be further leveraged to inform decisions in the bioenergy industry.

Broadly speaking, the contributions of this thesis are rooted in both technological and methodological advances. From the technological point of view, this thesis argues that traditional spatial decision support systems (SDSSs) (Densham 1991; Matthies et al. 2007), which are primarily designed and developed based on single desktop or server mode, are limited in terms of handling massive and various geographic data, solving computationally intensive decision support models, and allowing collaborative decision making. Therefore, a new SDSS framework powered by cyberGIS (Wang 2010; Wang et al. 2013; Wang 2017) is proposed in this thesis featuring a highly interactive online GIS user environment, seamless cyberinfrastructure access, and data and computing intensive spatial analysis methods in the context of agricultural and energy sustainability. A major benefit brought by the power of cyberGIS is to include uncertainty

quantification as part of the decision support system. Estimating uncertainty associated with decisions is often computationally intensive but important to decision makers, especially in complex decision support models. Through demonstrating a cyberGIS-powered SDSS namely CyberGIS-BioScope, this thesis reveals how the cyberGIS framework facilitates interactive decision support and computationally intensive uncertainty and sensitivity analysis on complex biomass supply chain optimization problems.

On the methodological side, this thesis examines two different approaches for uncertainty quantification. The first approach focuses on uncertainty propagation and global sensitivity analysis with the goal of understanding the variability of outcomes given model input uncertainty and quantifying the impact of each input that is responsible for this variability. This approach is often applied in cases when a decision model is available but input data are subject to uncertainty. In Chapter 3, this approach is applied to perform uncertainty and sensitivity analysis to quantify how various sources of uncertainty in the biomass supply chain contribute to the variation of optimal infrastructure configurations and bioethanol production cost. While the first approach emphasizes on how the uncertainty of data propagates through models, the second approach is more data-driven with a specific consideration of spatiotemporal features. The second approach aims to model the uncertainty of observed data by relating it to explanatory variables. More importantly, the relationship modeled in this approach could vary over space and time. In Chapter 4, a novel Bayesian hierarchical model is developed based on this approach to study the relationship between crop yields and weather dynamics. Both two uncertainty quantification approaches are demonstrated in the context of agriculture and energy sustainability case studies to help inform better decision support.

Future work will seek to further enhance the functionality as well as the performance of the cyberGIS-enabled spatial decision support systems (SDSS). On the technical side, more work could be done to improve the data interoperability and service integration. For example, the current system could be extended with capabilities to directly fetch data (biomass yield, weather) from online services such as USDA National Agricultural Statistics Service (https://www.nass.usda.gov/), or incorporate multiple cross-domain models to enable further understanding of the factors that influence the supply chain systems, including weather models (e.g., Weather Research Forecasting model, http://wrf-model.org) and multiple crop simulation models (e.g., BioCro, Miguez et al., 2012).

On the application side, future work could be done to further understand the complexity and sustainability of biomass-biofuel supply systems. In addition to cost-benefit analysis, multi-objective optimization with economic, environmental, and social measurements is desirable to be incorporated into the decision support system. Energy consumption, water footprint, and greenhouse gas (GHG) emissions are the three key environmental indicators representing the biomass–biofuels supply chain, from biomass production to transportation to biofuel conversion. It will be interesting to include environmental analysis as part of the decision model toward sustainable biofuel development. Another opportunity is to combine models that are at a finer scale of biomass supply chain management and implement operational level decision making. For example, optimal control of such processes as harvest scheduling, inventory planning, and transportation management can be incorporated into the current cyberGIS-based decision support system.

REFERENCES

ACIS, Applied Climate Information System, http://www.rcc-acis.org/index.html [accessed July 1, 2017]

Aerts, J. C., Goodchild, M. F., and Heuvelink, G. (2003). Accounting for spatial uncertainty in optimization with spatial decision support systems. *Transactions in GIS, 7*(2), 211-230.

An, L., Tsou, M. H., Crook, S. E., Chun, Y., Spitzberg, B., Gawron, J. M., and Gupta, D. K. (2015). Space–time analysis: Concepts, quantitative methods, and future directions. *Annals of the Association of American Geographers, 105*(5), 891-914.

Anselin, L. (2013). *Spatial econometrics: methods and models (Vol. 4)*. Springer Science and Business Media.

Ascough, J. C., Maier, H. R., Ravalico, J. K., and Strudley, M. W. (2008). Future research challenges for incorporation of uncertainty in environmental and ecological decision-making. *Ecological modelling*, *219*(3), 383-399.

Azaron, A., Brown, K. N., Tarim, S. A., and Modarres, M. (2008). A multi-objective stochastic programming approach for supply chain design considering risk. *International Journal of Production Economics, 116*(1), 129-138.

Balat, M., and Balat, H. (2009). Recent trends in global production and utilization of bio-ethanol fuel. *Applied energy, 86*(11), 2273-2282.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data.* Crc Press.

Barbarosoğlu, G., and Arda, Y. (2004). A two-stage stochastic programming framework for transportation planning in disaster response. *Journal of the operational research society*, *55*(1), 43-53.

Basile, R., Durbán, M., Mínguez, R., Montero, J. M., and Mur, J. (2014). Modeling regional economic dynamics: Spatial dependence, spatial heterogeneity and nonlinearities. *Journal of Economic Dynamics and Control, 48*, 229-245.

Basso, B., Cammarano, D., and Carfagna, E. (2013, July). Review of crop yield forecasting methods and early warning systems. In *Proceedings of the First Meeting of the Scientific Advisory Committee of the Global Strategy to Improve Agricultural and Rural Statistics, FAO Headquarters, Rome, Italy* (pp. 18-19).

Becher, S., and Kaltschmitt, M. (2013, May). Logistic Chains of Solid Biomass—Classification and Chain Analysis. In *Proceedings of the 8th European Biomass Conference* (pp. 401-408).

Ben-Tal, A., El Ghaoui, L., and Nemirovski, A. (2009). *Robust optimization*. Princeton University Press.

Besag, J. and Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika, 82*(4), 733-746.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, *43*(1), 1-20.

Birge, J. R., and Louveaux, F. (2011). *Introduction to stochastic programming*. Springer Science & Business Media.

Bornn, L. and Zidek, J. V. (2012). Efficient stabilization of crop yield prediction in the Canadian Prairies. *Agricultural and forest meteorology*, 152, 223-232.

Brechbill, S. C., Tyner, W. E., and Ileleji, K. E. (2011). The economics of biomass collection and transportation and its supply to Indiana cellulosic and electric utility facilities. *BioEnergy Research*, 4(2), 141-152.

Brunsdon, C., Fotheringham, A. S., and Charlton, M. E. (1996). Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical analysis, 28*(4), 281-298.

Brunsdon, C., Fotheringham, S., and Charlton, M. (1998). Geographically weighted regression. *Journal of the Royal Statistical Society: Series D (The Statistician), 47*(3), 431-443.

Brunsdon, C., Fotheringham, A. S., and Charlton, M. (1998). Spatial nonstationarity and autoregressive models. *Environment and Planning A, 30*(6), 957-973.

Cai, R., Yu, D., and Oppenheimer, M. (2014). Estimating the Spatially Varying Responses of Corn Yields to Weather Variations using Geographically Weighted Panel Regression. *Journal of Agricultural and Resource Economics, 39*(2).

Chakhar, S., and Mousseau, V. (2003, October). Towards a typology of spatial decision problems. In *The 58th Meeting of the European Working Group Multiple Criteria Decision Aiding, Moscow, Russia, October 9-11*. EWG MCDA.

Chatfield, C. (2006). Model uncertainty. *Encyclopedia of Environmetrics*.

Choi, J., Lawson, A. B., Cai, B., Hossain, M. M., Kirby, R. S., and Liu, J. (2012). A Bayesian latent model with spatio-temporally varying coefficients in low birth weight incidence data. *Statistical methods in medical research, 21*(5), 445-456.

Church, R. L. (2001). Spatial optimization models. In *International encyclopedia of the social & behavioral sciences,* ed. N. J. Smelser and P. B. Baltes, 811–18. New York: Elsevier.

Church, R. L., and Cova, T. J. (2000). Mapping evacuation risk on transportation networks using a spatial optimization model. *Transportation Research Part C: Emerging Technologies, 8*(1), 321-336.

Church, R. L., and Murray, A. T. (2009). *Business site selection, location analysis and GIS*. New York: Wiley

Cressie, N., Calder, C. A., Clark, J. S., Hoef, J. M. V., and Wikle, C. K. (2009). Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. *Ecological Applications*, *19*(3), 553-570.

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

Crosetto, M., and Tarantola, S. (2001). Uncertainty and sensitivity analysis: tools for GIS-based model implementation. *International Journal of Geographical Information Science*, *15*(5), 415-437.

Cukier, R. I., Levine, H. B., and Shuler, K. E. (1978). Nonlinear sensitivity analysis of multiparameter model systems. *Journal of computational physics*, *26*(1), 1-42.

Densham, P. J. (1991). Spatial decision support systems. *Geographical information systems: Principles and applications*, 1, 403-412.

Dong, G., Ma, J., Harris, R., and Pryce, G. (2016). Spatial random slope multilevel modeling using multivariate conditional autoregressive models: A case study of subjective travel satisfaction in Beijing. *Annals of the American Association of Geographers*, *106*(1), 19-35.

Efron, B., and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Feizizadeh, B., Jankowski, P., and Blaschke, T. (2014). A GIS based spatially-explicit sensitivity and uncertainty analysis approach for multi-criteria decision analysis. *Computers & geosciences*, *64*, 81-95.

Finley, A. O. (2011). Comparing spatially-varying coefficients models for analysis of ecological data with non-stationary and anisotropic residual dependence. *Methods in Ecology and Evolution, 2*(2), 143-154.

Fiorese, G., and Guariso, G. (2010). A GIS-based approach to evaluate biomass potential from energy crops at regional scale. *Environmental Modelling & Software*, *25*(6), 702-711.

Fisher, P. F. (1999). Models of uncertainty in spatial data. *Geographical information systems*, 1, 191-205.

Fotheringham, A. S., Charlton, M. E., and Brunsdon, C. (1998). Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis. *Environment and planning A*, *30*(11), 1905-1927.

Fotheringham, A. S., Brunsdon, C., and Charlton, M. (2002). *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons.

Fotheringham, A. S., Crespo, R., and Yao, J. (2015). Geographical and temporal weighted regression (GTWR). *Geographical Analysis*, *47*(4), 431-452.

Freppaz, D., Minciardi, R., Robba, M., Rovatti, M., Sacile, R., and Taramasso, A. (2004). Optimizing forest biomass exploitation for energy supply at a regional level. B*iomass and Bioenergy*, *26*(1), 15-25.

Frombo, F., Minciardi, R., Robba, M., and Sacile, R. (2009). A decision support system for planning biomass-based energy production. *Energy*, *34*(3), 362-369.

Gahegan, M., and Ehlers, M. (2000). A framework for the modelling of uncertainty between remote sensing and geographic information systems.*ISPRS Journal of Photogrammetry and Remote Sensing*, *55*(3), 176-188.

Gelfand, A. E., Kim, H. J., Sirmans, C. F., and Banerjee, S. (2003). Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, *98*(462), 387-396.

Gewin, V. (2004). Mapping opportunities. *Nature*, *427*(6972), 376-377.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. (Eds.). (1995). *Markov chain Monte Carlo in practice.* CRC press.

Gómez-Delgado, M., and Tarantola, S. (2006). GLOBAL sensitivity analysis, GIS and multi-criteria evaluation for a sustainable planning of a hazardous waste disposal site in Spain. *International Journal of Geographical Information Science, 20*(4), 449-466.

Graham, R. L., English, B. C., and Noon, C. E. (2000). A geographic information system-based modeling system for evaluating the cost of delivered energy crop feedstock. *Biomass and bioenergy, 18*(4), 309-329.

Griffith, D. A. (2008). Spatial-filtering-based contributions to a critique of geographically weighted regression (GWR). *Environment and Planning A, 40*(11), 2751-2769.

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

Hess, J. R., Wright, C. T., and Kenney, K. L. (2007). Cellulosic biomass feedstocks and logistics for ethanol production. *Biofuels, Bioproducts and Biorefining, 1*(3), 181-190.

Heuvelink, G. B., Burrough, P. A., and Stein, A. (1989). Propagation of errors in spatial modelling with GIS. *International Journal of Geographical Information System, 3*(4), 303-322.

Hoeting, J. A. (2009). The importance of accounting for spatial and temporal correlation in analyses of ecological data. *Ecological Applications, 19*(3), 574-577.

Höhn, J., Lehtonen, E., Rasi, S., and Rintala, J. (2014). A Geographical Information System (GIS) based methodology for determination of potential biomasses and sites for biogas plants in southern Finland. *Applied Energy, 113*, 1-10.

Homma, T., and Saltelli, A. (1996). Importance measures in global sensitivity analysis of nonlinear models. *Reliability Engineering & System Safety, 52*(1), 1-17.

Hsiao, C. (2014). *Analysis of panel data* (No. 54). Cambridge university press.

Hu, H., Lin, T., Liu, Y. Y., Wang, S., and Rodríguez, L. F. (2015). CyberGIS-BioScope: a cyberinfrastructure-based spatial decision-making environment for biomass-to-biofuel supply chain optimization. *Concurrency and Computation: Practice and Experience*, *27*(16), 4437-4450.

Huang, B., Wu, B., and Barry, M. (2010). Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices. *International Journal of Geographical Information Science, 24*(3), 383-401.

Huang, Y., Chen, C. W., and Fan, Y. (2010). Multistage optimization of the supply chains of biofuels. *Transportation Research Part E: Logistics and Transportation Review*, 46(6), 820-830.

Humbird, D., Davis, R., Tao, L., Kinchin, C., Hsu, D., Aden, A., Schoen, P., Lukas, J., Olthof, B., Worley, M. and Sexton, D. (2011). *Process design and economics for biochemical conversion of lignocellulosic biomass to ethanol: dilute-acid pretreatment and enzymatic hydrolysis of corn stover.* (No. NREL/TP-5100-47764). National Renewable Energy Laboratory (NREL), Golden, CO.

Hunter, G. J., and Goodchild, M. F. (1993, July). Managing uncertainty in spatial databases: Putting theory into practice. In *Papers from the Annual Conference, URISA: Urban And Regional Information Systems* (pp. 1-14), Atlanta, July 25-29.

Jain, A. K., Khanna, M., Erickson, M., and Huang, H. (2010). An integrated biogeochemical and economic analysis of bioenergy crops in the Midwestern United States. *Global Change Biology - Bioenergy*, *2*(5), 217-234.

Jankowski, P., Nyerges, T. L., Smith, A., Moore, T. J., and Horvath, E. (1997). Spatial group choice: a SDSS tool for collaborative spatial decision making. *International journal of geographical information science*, 11(6), 577-602.

Jansen, M. J. (1999). Analysis of variance designs for model output. *Computer Physics Communications*, *117*(1), 35-43.

Jiang, P., He, Z., Kitchen, N. R., and Sudduth, K. A. (2009). Bayesian analysis of within-field variability of corn yield using a spatial hierarchical model. *Precision Agriculture*, *10*(2), 111-127.

Jones, J. W., Hoogenboom, G., Porter, C. H., Boote, K. J., Batchelor, W. D., Hunt, L. A., ... and Ritchie, J. T. (2003). The DSSAT cropping system model. *European journal of agronomy*, *18*(3), 235-265.

Jonker, J. G. G., Junginger, H. M., Verstegen, J. A., Lin, T., Rodríguez, L. F., Ting, K. C., A. P. C. Faaij, and F. van der Hilst. (2016). Supply chain optimization of sugarcane first generation and eucalyptus second generation ethanol production in Brazil. *Applied Energy*, *173*, 494-510.

Keller, C. P. (1990). Decision support multiple criteria methods. *NCGIA Core Curriculum, National Center for Geographic Information and Analysis*.

Kim, J., Realff, M. J., and Lee, J. H. (2011). Optimal design and global sensitivity analysis of biomass supply chain networks for biofuels under uncertainty. *Computers & Chemical Engineering*, *35*(9), 1738-1751.

Kleywegt, A. J., Shapiro, A., and Homem-de-Mello, T. (2002). The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, *12*(2), 479-502.

Kucherenko, S. S. (2005). On global sensitivity analysis of quasi-Monte Carlo algorithms. *Monte Carlo Methods and Applications*, *11*(1), 83-92.

Kwan, M. P. (2016). Algorithmic geographies: Big data, algorithmic uncertainty, and the production of geographic knowledge. *Annals of the American Association of Geographers*, *106*(2), 274-282.

Leung, Y. (2012). *Intelligent spatial decision support systems*. Springer Science & Business Media.

Lilburne, L., and Tarantola, S. (2009). Sensitivity analysis of spatial models. *International Journal of Geographical Information Science*, *23*(2), 151-168.

Ligmann-Zielinska, A., Church, R. L., and Jankowski, P. (2008). Spatial optimization as a generative technique for sustainable multiobjective land-use allocation. *International Journal of Geographical Information Science*, *22*(6), 601-622.

Ligmann-Zielinska, A., and Jankowski, P. (2014). Spatially-explicit integrated uncertainty and sensitivity analysis of criteria weights in multicriteria land suitability evaluation. *Environmental Modelling & Software*, *57*, 235-247.

Lin, T., Rodríguez, L. F., Shastri, Y. N., Hansen, A. C., and Ting, K. C. (2013). GIS-enabled biomass-ethanol supply chain optimization: model development and Miscanthus application. *Biofuels, Bioproducts and Biorefining*, *7*(3), 314-333.

Lin, T., Rodríguez, L. F., Shastri, Y. N., Hansen, A. C., and Ting, K. C. (2014). Integrated strategic and tactical biomass–biofuel supply chain optimization. *Bioresource technology*, *156*, 256-266.

Liu, Y., Padmanabhan, A., and Wang, S. (2015). CyberGIS Gateway for enabling data-rich geospatial research and education. *Concurrency and Computation: Practice and Experience*, *27*(2), 395-407.

Lin, T., Wang, S., Rodríguez, L. F., Hu, H., and Liu, Y. (2015). CyberGIS-enabled decision support platform for biomass supply chain optimization. *Environmental Modelling & Software, 70*, 138-148.

Liu, Y. Y., and Wang, S. (2015). A scalable parallel genetic algorithm for the generalized assignment problem. *Parallel computing*, *46*, 98-119.

Lobell, D. B. and Burke, M. B. (2010). On the use of statistical models to predict crop yield responses to climate change. *Agricultural and Forest Meteorology, 150*(11), 1443-1452.

Longley, P. A., Goodchild, M. F., Maguire, D. J., and Rhind, D. W. (2001).*Geographic information systems and science*. John Wiley & Sons Ltd.

Luxen, D., and Vetter, C. (2011, November). Real-time routing with OpenStreetMap data. In *Proceedings of the 19th ACM SIGSPATIAL international conference on advances in geographic information systems* (pp. 513-516). ACM.

Malczewski, J. (1999). *GIS and multicriteria decision analysis*. John Wiley & Sons.

Marufuzzaman, M., Eksioglu, S. D., and Huang, Y. E. (2014). Two-stage stochastic programming supply chain model for biodiesel production via wastewater treatment. *Computers & Operations Research*, *49*, 1-17.

Marvin, W. A., Schmidt, L. D., Benjaafar, S., Tiffany, D. G., and Daoutidis, P. (2012). Economic optimization of a lignocellulosic biomass-to-ethanol supply chain. *Chemical Engineering Science*, 67(1), 68-79.

Matthies, M., Giupponi, C., and Ostendorf, B. (2007). Environmental decision support systems: Current issues, methods and tools. *Environmental Modelling & Software*, *22*(2), 123-127.

McGrath, J. M., Betzelberger, A. M., Wang, S., Shook, E., Zhu, X., Long, S. P., and Ainsworth, E. (2015) An Analysis of Ozone Damage to Historical Maize and Soybean Yields in the United States". *Proceedings of the National Academy of Sciences, 112*(46): 14390-14395.

McMaster, G. S. and Wilhelm, W. W. (1997). Growing degree-days: one equation, two interpretations. *Agricultural and forest meteorology*, *87*(4), 291-300.

Melo, M. T., Nickel, S., and Saldanha-da-Gama, F. (2009). Facility location and supply chain management–A review. *European journal of operational research*, *196*(2), 401-412.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, *37*(1/2), 17-23.

Munda, G. (2012). *Multicriteria evaluation in a fuzzy environment: theory and applications in ecological economics*. Springer Science & Business Media.

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

Nagel, J. (2000). Determination of an economic energy supply structure based on biomass using a mixed-integer linear optimization model. *Ecological Engineering*, *16*, 91-102.

Noon, C. E., and Daly, M. J. (1996). GIS-based biomass resource assessment with BRAVO. *Biomass and Bioenergy*, *10*(2-3), 101-109.

Olesen, J. E., Trnka, M., Kersebaum, K. C., Skjelvåg, A. O., Seguin, B., Peltonen-Sainio, P., ... and Micale, F. (2011). Impacts and adaptation of European crop production systems to climate change. *European Journal of Agronomy*, *34*(2), 96-112.

Ozaki, V. A., Ghosh, S. K., Goodwin, B. K., and Shirota, R. (2008). Spatio-temporal modeling of agricultural yield data with an application to pricing crop insurance contracts. *American Journal of Agricultural Economics, 90*(4), 951-961.

Panichelli, L., and Gnansounou, E. (2008). GIS-based approach for defining bioenergy facilities location: A case study in Northern Spain based on marginal delivery costs and resources competition between facilities. *Biomass and Bioenergy, 32*(4), 289-300.

Parent, E., and Rivot, E. (2012). *Introduction to hierarchical Bayesian modeling for ecological data*. CRC Press.

Parker, N., Fan, Y., and Ogden, J. (2010). From waste to hydrogen: an optimal design of energy production and distribution network. *Transportation Research Part E: Logistics and Transportation Review, 46*(4), 534-545.

Perimenis, A., Walimwipi, H., Zinoviev, S., Müller-Langer, F., and Miertus, S. (2011). Development of a decision support tool for the assessment of biofuels. *Energy Policy, 39*(3), 1782-1793.

Pipkin, J. S. (1991). Spatial Analysis and Planning under Imprecision, by Y. Leung. *Geographical Analysis*, *23*(1), 90-92.

Pishvaee, M. S., Rabbani, M., and Torabi, S. A. (2011). A robust optimization approach to closed-loop supply chain network design under uncertainty.*Applied Mathematical Modelling*, *35*(2), 637-649.

PRISM Climate Group, 2004, Oregon State University, http://prism.oregonstate.edu [accessed July 1, 2017].

Ray, D. K., Gerber, J. S., MacDonald, G. K., and West, P. C. (2015). Climate variation explains a third of global crop yield variability. *Nature Communications*, 6.

Refsgaard, J. C., van der Sluijs, J. P., Højberg, A. L., and Vanrolleghem, P. A. (2007). Uncertainty in the environmental modelling process–a framework and guidance. *Environmental modelling & software*, *22*(11), 1543-1556.

Rentizelas, A. A., Tatsiopoulos, I. P., and Tolis, A. (2009). An optimization model for multi-biomass tri-generation energy supply. *Biomass and bioenergy*, *33*(2), 223-233.

Saltelli, A., Tarantola, S., and Chan, K. S. (1999). A quantitative model-independent method for global sensitivity analysis of model output.*Technometrics*, *41*(1), 39-56.

Saltelli, A., Chan, K., and Scott, E. M. (2000). Sensitivity analysis Wiley series in probability and statistics. *Willey, New York*.

Saltelli, A. (2002). Making best use of model evaluations to compute sensitivity indices. *Computer Physics Communications*, *145*(2), 280-297.

Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., and Tarantola, S. (2010). Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. *Computer Physics Communications*,*181*(2), 259-270.

Santoso, T., Ahmed, S., Goetschalckx, M., and Shapiro, A. (2005). A stochastic programming approach for supply chain network design under uncertainty. *European Journal of Operational Research*, *167*(1), 96-115.

Schlenker, W. and Roberts, M. J. (2009). Nonlinear temperature effects indicate severe damages to US crop yields under climate change. *Proceedings of the National Academy of sciences*, *106*(37), 15594-15598.

Shapiro, A., and Philpott, A. (2007). A tutorial on stochastic programming.*Manuscript. Available at www2. isye. gatech. edu/ashapiro/publications. html*, *17*.

Shapiro, J. (2006). *Modeling the supply chain*. Nelson Education.

Sharma, B., Ingalls, R. G., Jones, C. L., and Khanchi, A. (2013). Biomass supply chain design and analysis: Basis, overview, modeling, challenges, and future. *Renewable and Sustainable Energy Reviews*, *24*, 608-627.

Sharma, V., Irmak, A., Kabenge, I., and Irmak, S. (2011). Application of GIS and geographically weighted regression to evaluate the spatial non-stationarity relationships between precipitation vs. irrigated and rainfed maize and soybean yields. *Transactions of the ASABE*, *54*(3), 953-972.

Shi, W., Fisher, P., and Goodchild, M. F. (2003). *Spatial data quality*. CRC Press.

Shi, W. (2009). *Principles of modeling uncertainties in spatial data and spatial analyses*. CRC Press.

Sobol, I. Y. M. (1990). On sensitivity estimation for nonlinear mathematical models. *Matematicheskoe Modelirovanie*, *2*(1), 112-118.

Sobol, I. M. (2001). Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and computers in simulation, 55*(1), 271-280.

Sobol, I. M., and Kucherenko, S. S. (2005). On global sensitivity analysis of quasi-Monte Carlo algorithms. *Monte Carlo Methods and Applications*, *11*(1), 83-92.

Sodhi, M. S., and Tang, C. S. (2009). Modeling supply-chain planning under demand uncertainty using stochastic programming: A survey motivated by asset–liability management. *International Journal of Production Economics,121*(2), 728-738.

Srirangan, K., Akawi, L., Moo-Young, M., and Chou, C. P. (2012). Towards sustainable production of clean energy carriers from biomass resources. *Applied energy, 100*, 172-186.

Steinitz, C. (2012). *A framework for geodesign: Changing geography by design.* Redlands, CA: esri.

Tíba, C., Candeias, A. L. B., Fraidenraich, N., Barbosa, E. D. S., de Carvalho Neto, P. B., and de Melo Filho, J. B. (2010). A GIS-based decision support tool for renewable energy management and planning in semi-arid rural environments of northeast of Brazil. *Renewable Energy*, *35*(12), 2921-2932.

Tilman, D., Socolow, R., Foley, J.A., Hill, J., Larson, E., Lynd, L., Pacala, S., Reilly, J., Searchinger, T., Somerville, C. and Williams, R. (2009). Beneficial biofuels—the food, energy, and environment trilemma. *Science*, *325*(5938), 270-271.

Tong, D., and Murray, A. T. (2012). Spatial optimization in geography. *Annals of the Association of American Geographers*, *102*(6), 1290-1309.

Tyndall, J. C., Berg, E. J., and Colletti, J. P. (2011). Corn stover as a biofuel feedstock in Iowa's bio-economy: an Iowa farmer survey. *Biomass and Bioenergy*, *35*(4), 1485-1495.

USDA. The Census of Agriculture. USDA National Agricultural Statistics Service https://quickstats.nass.usda.gov/ [accessed July 1, 2017].

Uusitalo, L., Lehikoinen, A., Helle, I., and Myrberg, K. (2015). An overview of methods to evaluate uncertainty of deterministic models in decision support. *Environmental Modelling & Software*, *63*, 24-31.

Vlachos, D., Iakovou, E., Karagiannidis, A., and Toka, A. (2008, October). A strategic supply chain management model for waste biomass networks. In *3rd International Conference on Manufacturing Engineering* (pp. 797-804).

Voivontas, D., Assimacopoulos, D., and Koukios, E. G. (2001). Aessessment of biomass potential for power production: a GIS based method. *Biomass and bioenergy*, *20*(2), 101-112.

Waller, L. A., Zhu, L., Gotway, C. A., Gorman, D. M., and Gruenewald, P. J. (2007). Quantifying geographic variations in associations between alcohol distribution and violence: a

comparison of geographically weighted regression and spatially varying coefficient models. *Stochastic Environmental Research and Risk Assessment, 21*(5), 573-588.

Wang, S., and Liu, Y. (2009). TeraGrid GIScience gateway: bridging cyberinfrastructure and GIScience. *International Journal of Geographical Information Science, 23*(5), 631-656.

Wang, S. (2010). A CyberGIS framework for the synthesis of cyberinfrastructure, GIS, and spatial analysis. *Annals of the Association of American Geographers*, *100*(3), 535-557.

Wang, S., Anselin, L., Bhaduri, B., Crosby, C., Goodchild, M. F., Liu, Y., and Nyerges, T. L. (2013). CyberGIS software: a synthetic review and integration roadmap. *International Journal of Geographical Information Science, 27*(11), 2122-2145.

Wang, S. CyberGIS. *The International Encyclopedia of Geography*. John Wiley & Sons, Ltd; 2017. doi:10.1002/9781118786352.wbieg0931.

Wheeler, D. C. and Calder, C. A. (2007). An assessment of coefficient accuracy in linear regression models with spatially varying coefficients. *Journal of Geographical Systems, 9*(2), 145-166.

Wheeler, D. C. and Waller, L. A. (2009). Comparing spatially varying coefficient models: a case study examining violent crime rates and their relationships to alcohol outlets and illegal drug arrests. *Journal of Geographical Systems, 11*(1), 1-22.

Williams, H. P. (2009). *Logic and integer programming* (Vol. 130). Berlin: Springer.

Wu, B., Li, R., and Huang, B. (2014). A geographically and temporally weighted autoregressive model with application to housing prices. *International Journal of Geographical Information Science, 28*(5), 1186-1204.

Wyman, C. E. (2007). What is (and is not) vital to advancing cellulosic ethanol. *TRENDS in Biotechnology*, *25*(4), 153-157.

Xiao, N. (2008). A unified conceptual framework for geographical optimization using evolutionary algorithms. *Annals of the Association of American Geographers*, *98*(4), 795-817.

Yadav, Y. S., and Yadav, Y. K. (2016). Biomass Supply Chain Management: Perspectives and Challenges. In *Proceedings of the First International Conference on Recent Advances in Bioenergy Research* (pp. 267-281). Springer, New Delhi.

You, F., and Wang, B. (2011). Life cycle optimization of biomass-to-liquid supply chains with distributed–centralized processing networks. *Industrial & Engineering Chemistry Research*, *50*(17), 10102-10127.

You, F., Tao, L., Graziano, D. J., and Snyder, S. W. (2012). Optimal design of sustainable cellulosic biofuel supply chains: multiobjective optimization coupled with life cycle assessment and input–output analysis. *AIChE Journal*, *58*(4), 1157-1180.

Zambelli, P., Lora, C., Spinelli, R., Tattoni, C., Vitti, A., Zatelli, P., and Ciolli, M. (2012). A GIS decision support system for regional forest management to assess biomass availability for renewable energy production. *Environmental Modelling & Software*, *38*, 203-213.

Zhang, J., and Goodchild, M. F. (2002). *Uncertainty in geographical information*. CRC press.

Zhang, J., Osmani, A., Awudu, I., and Gonela, V. (2013). An integrated optimization model for switchgrass-based bioethanol supply chain. *Applied Energy, 102*, 1205-121

APPENDIX A

Construction of matrices A, B and $A_B^{(i)}$, and $B_A^{(i)}$ from Sobol Quasi-random Sequence

$\boldsymbol{A}$ and $\boldsymbol{B}$ (Fig. A.1 top two) are matrices formed by two independent samples of model inputs. Both matrices are in size $N \times k$, where N is number of simulation for each matrix and k is the number of input parameters. Construction of $\boldsymbol{A}$ and $\boldsymbol{B}$ are based on Monte Carlo sampling methods with a quasi-random sequence. The reason for using quasi-random sequences such as Sobol sequence instead of crude Monte Carlo sampling is that it has faster rate of convergence in the estimation of multi-dimensional integrals [36]. An example of Sobol quasi-random sequence with N=8 and k=8 from a 0 to 1 uniform distribution is shown in Table A.1.

Table A.1 First eight points in an eight-dimensional Sobol quasi-random sequence

0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000
0.7500	0.2500	0.7500	0.2500	0.7500	0.2500	0.7500	0.2500
0.2500	0.7500	0.2500	0.7500	0.2500	0.7500	0.2500	0.7500
0.3750	0.3750	0.6250	0.1250	0.8750	0.8750	0.1250	0.6250
0.8750	0.8750	0.1250	0.6250	0.3750	0.3750	0.6250	0.1250
0.6250	0.1250	0.3750	0.3750	0.1250	0.6250	0.8750	0.8750
0.1250	0.6250	0.8750	0.8750	0.6250	0.1250	0.3750	0.3750
0.1875	0.3125	0.3125	0.6875	0.5625	0.1875	0.0625	0.9375

Based on Sobol's sequences, matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ of size (N, k) can be generated from a quasi-random sequence of size (N, 2k) with $\boldsymbol{A}$ from the left half and $\boldsymbol{B}$ from the right part. Then matrix

$\boldsymbol{A}_{\boldsymbol{B}}^{(i)}$ is defined as all columns of $\boldsymbol{A}$, except the *i*th column taken from $\boldsymbol{B}$, and matrix $\boldsymbol{B}_{\boldsymbol{A}}^{(i)}$ formed by the *i*th column of $\boldsymbol{A}$ and all the remaining columns of from $\boldsymbol{B}$ (Fig. A.1).

$$\boldsymbol{A} = \begin{bmatrix} x_{a1}^{(1)} & \cdots & x_{a1}^{(i)} & \cdots & x_{a1}^{(k)} \\ x_{a2}^{(1)} & \cdots & x_{a2}^{(i)} & \cdots & x_{a2}^{(k)} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ x_{aN}^{(1)} & \cdots & x_{aN}^{(i)} & \cdots & x_{aN}^{(k)} \end{bmatrix} \quad \boldsymbol{B} = \begin{bmatrix} x_{b1}^{(1)} & \cdots & x_{b1}^{(i)} & \cdots & x_{b1}^{(k)} \\ x_{b2}^{(1)} & \cdots & x_{b2}^{(i)} & \cdots & x_{b2}^{(k)} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ x_{bN}^{(1)} & \cdots & x_{bN}^{(i)} & \cdots & x_{bN}^{(k)} \end{bmatrix}$$

$$\boldsymbol{A}_{\boldsymbol{B}}^{(\boldsymbol{i})} = \begin{bmatrix} x_{a1}^{(1)} & \cdots & x_{b1}^{(i)} & \cdots & x_{a1}^{(k)} \\ x_{a2}^{(1)} & \cdots & x_{b2}^{(i)} & \cdots & x_{a2}^{(k)} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ x_{aN}^{(1)} & \cdots & x_{bN}^{(i)} & \cdots & x_{aN}^{(k)} \end{bmatrix} \boldsymbol{B}_{\boldsymbol{A}}^{(\boldsymbol{i})} = \begin{bmatrix} x_{b1}^{(1)} & \cdots & x_{a1}^{(i)} & \cdots & x_{b1}^{(k)} \\ x_{b2}^{(1)} & \cdots & x_{a2}^{(i)} & \cdots & x_{b2}^{(k)} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ x_{bN}^{(1)} & \cdots & x_{aN}^{(i)} & \cdots & x_{bN}^{(k)} \end{bmatrix}$$

Fig. A.1 – Generating matrices $\boldsymbol{A}_{\boldsymbol{B}}^{(\boldsymbol{i})}$ and $\boldsymbol{B}_{\boldsymbol{A}}^{(\boldsymbol{i})}$ from $\boldsymbol{A}$ and $\boldsymbol{B}$ [20]

APPENDIX B

Nomenclature for Stochastic BioScope Model

Table B.1 A list of decision variables

Symbol	Type		Description
C_B^s	Non-negative variable	continuous	Biomass procurement costs of scenario *s*
C_M^s	Non-negative variable	continuous	Biomass transportation costs of scenario *s*
C_S	Non-negative variable	continuous	CSP related costs
C_E	Non-negative variable	continuous	Biorefinery related costs
$f^{i,j,s}$	Non-negative variable	continuous	Amount of biomass flow from supply to CSP of scenario *s*
$f^{j,k,s}$	Non-negative variable	continuous	Amount of biomass flow from CSP to biorefinery of scenario *s*
p^j	Non-negative variable	continuous	The total centralized storage and preprocessing (CSP) facility capacity in county *j*
o_s^j	Binary variable		Indicates whether there is a CSP facility located in county *j*

Table B.1 (cont.).

$p^{j,l}$	Non-negative variable	continuous	The CSP facility capacity in county j at level l
$o_s^{j,l}$	Binary variable		Indicates whether there is a CSP facility located in county j at level l
q^k	Non-negative variable	continuous	The total biorefinery capacity in county k
o_s^k	Binary variable		Indicates whether there is a biorefinery facility located in county k
$q^{k,l}$	Non-negative variable	continuous	The biorefinery capacity in county k at level l
$o_e^{k,l}$	Binary variable		Indicate whether there is a biorefinery facility located in county k at level l

Table B.2 A list of model input data and parameters

Symbol	Description
ρ^s	Probability density of scenario s
$b^{i,s}$	County-level biomass availability of scenario s
$c^{i,s}$	County-level biomass purchase cost of scenario s
Q	Total biomass required for processing

Table B.2 (cont.).

$d^{i,j}$	Distance between biomass supply sites and centralized storage and preprocessing (CSP) sites
$d^{j,k}$	Distance between CSP sites and biorefinery sites
s_{op}	Unit operating costs for CSP
s_v^l	Variable capital costs for CSP at different levels
s_f^l	Fixed capital costs for CSP at different levels
α	Annualized cost factor
β	Biomass loss rate at CSP
e_{op}	Unit operating costs for a biorefinery
e_v^l	Variable capital costs for a biorefinery at different levels
e_f^l	Fixed capital costs for a biorefinery at different levels