AGENT-BASED MODELS TO COUPLE NATURAL AND HUMAN SYSTEMS FOR WATERSHED MANAGEMENT ANALYSIS

BY

YAO HU

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Civil Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2016

Urbana, Illinois

Doctoral Committee:

Professor Ximing Cai, Chair
Professor Albert J. Valocchi, Co-Chair
Professor Shaowen Wang
Associate Professor Negar Kiyavash
Associate Professor Nicolas Brozović, University of Nebraska
Assistant Professor Christopher Quinn, Purdue University

ABSTRACT

This dissertation expands conventional physically-based environmental models with human factors for watershed management analysis. Using an agent-based modeling framework, two approaches, one based on optimization and the other on data mining-are applied to modeling farmers' pumping decision-making processes in the High Plains aquifer within the hydrological observatory area. The resulting agent-based models (ABMs) are coupled with a physically-based groundwater model to investigate the interactions between farmers and the underlying groundwater system.

With the optimization-based approach, the computational intensity arises from the execution of the resulting coupled ABM and groundwater model. This dissertation develops a computational framework that utilizes multithreaded programming and Hadoop-based cloud computing to address the computational issues. The framework allows multiple users to access and execute the web-based application of the coupled models simultaneously without an increase in latency via computer network. In addition, another computational framework to combine Hadoop-based Cloud Computing techniques with Polynomial Chaos Expansion (PCE) based variance decomposition approach is developed to conduct global sensitivity analysis with the coupled models, and influential behavioral parameters which are used to simulate agents' behavior are identified.

Being different from the optimization-based approach, which assumes all agents are rational, the data-driven approach attempts to account for the influences of agents' bounded rationality on their behavior. A directed information graph (DIG) algorithm is used to exploit the causal relationships between agents' decisions (i.e., groundwater irrigation depth) and time-series of environmental, socio-economical and institutional variables, and a machine learning technique, boosted regression tree (BRT) is applied to converting these causal relationships to agents' behavioral rules. It is found that, in comparison with the optimization-based approach, crop profits and water tables as the result of agents' pumping behavior derived using the data-driven approach can better mimic the actual observations. Thus, we can conclude that the data-driven approach using DIG and BRT outperforms the optimization-based approach when capturing agents' pumping behavioral uncertainty as the result of bounded rationality, and for simulating real-world behaviors of agents.

摘要

本文通过将人为因素整合于物理环境模型中的方式， 对传统的流域管理和分析模型进行了扩展。 基于人工智能体模型的框架， 我们开发了两种方法来模拟处于 High Plains 含水层水文观测站区域内农民抽取地下水的决策过程； 其中一种方式是基于优化的方法，另一种是基于数据挖掘的方法。基于这两种方法所得到的人工智能体模型分别于基于物理过程的地下水模型进行耦合。这些耦合后的模型将用于研究农民和地下水系统的互动行为。

基于优化方法开发的耦合人工智能体和地下水模型的计算强度高。为了解决运算强度大的问题，本文作者开发了一种基于多线程和云计算(Hadoop)的框架。 在这个框架下，多个用户在不经历延时的情况下可通过用户界面同时调用在服务器端的耦合模型。此外，结合云计算(Hadoop)和多项式混沌展开(PCE)的方差分解方法，作者开发了一个用于耦合模型的全局性敏感分析的框架。通过这个框架，我们可以识别模拟智能体行为中具有影响力的行为参数。

采用优化方法设计的智能体，每个智能体都被假设为理性决策者。 与优化方法不同的是，采用数据驱动方式设计智能体时，会考虑到智能体的限制理性对于智能体行为的影响。一种被称为有向信息图 (DIG) 的算法被用于去发掘智能体的决策 (i.e., 地下水灌溉深度) 与相关的环境，社会经济和体制因素的因果关系。一种基于机器学习的增强性回归树 (BRT) 方法将上述因果关系转换为智能体的行为规则。相对于优化方法， 通过数据驱动方式得出的行为规则能够更好的模拟实际观察到的稻谷价格和地下水水位。由此我们可以得出， 使用有向信息图和增强性回归树的数据驱动方式能够更有效的捕获到因限制理性所带来的行为不确定性，从而更好的去模拟现实生活中智能体的行为。

To my grandparents and parents

ACKNOWLEDGMENTS

The PhD experience to me at University of Illinois is the mixture of anxiety and excitement, pain and joy. I still remember when I firstly came to U of I at year 2010 right after I finished my post-graduate work at Helmholtz Centre for Environmental Research (UFZ, Leipzig, Germany), I was very excited to start the new era of my life and explore new things. However, the excitement and curiosity were quickly replaced with the intensive coursework, peer pressure, adviser's expectation and my own ambition to do good research in a new and interdisciplinary area. Thanks to a group of friends and mentors outside of academia, in particular Bo Mu, Natalie Mullen and Steve Choi, I am able to make a soft landing.

With a year of exploring the new topic, agent-based model, I realize its potentials and become fascinated with it. I want to thank my adviser, Dr. Ximing Cai who brought me into the field, and the immense support, guidance and freedom he gave to me to explore and research. I also want to thank my other committee members, Dr. Albert Valocchi (Co-Adviser), Dr. Nicholas Brozović, Dr. Negar Kiyavash, Dr. Christopher Quinn and Dr. Shaowen Wang for their questions and comments, which help me think deeper and more critical about my research. Dr. Albert Valocchi and Dr. Nicholas Brozović have been great mentors to me. I really appreciate their personal guidance. In addition, I want to express my thanks to my former adviser Dr. Anke Hildebrandt and my beloved colleagues and friends at UFZ, such as Dr. Christoph Schneider and Dr. Falk Hesse. They help me become a critical thinker and an independent researcher.

To me, the most wonderful thing to study at the University of Illinois at Urbana-Champaign is to have the unique opportunity to explore various disciplines and bridge the knowledge gaps. While I was exploring ideas from different domains in an attempt to narrow down my research directions, I am very fortunate to know some colleagues and friends during my PhD. Benjamin DuPont and Oscar Garcia, without your trust and dedications to my research ideas, I would not have the courage and confidence to continue on the path. Noah Garfinkle and Dr. Christopher Quinn, with your tremendous help with my writing and countless hours of discussions we have, I am able to make my research work solid and publishable.

As a PhD student, like others I dedicate nearly all my available time to my work. Obviously, my work and life is not balanced. However, life could be much harder to me without my dear friends. They make me realize life is not only about science. Here, I want to thank Marianne Boix, Tru Luv Flo, Erhu Du, Oscar Garcia, Noah Garfinkle, Susana Kimura, Zheng Li, Aihui Ma, Christopher Quinn, Pascal Rademacher and Rui Zhuang. They are the sources of my comfort and energy.

Without the immense support from and sacrifice of my family, to get a PhD degree in U.S. could only be the story of others. My parents swallow their worries and pains that I am so far away from them, and my younger brother takes the solo responsibility to take care of our parents and grandparents, which enable me to concentrate on my study. This is the love that I am so lucky to have.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 CHAPTER I - INTRODUCTION

1.1 Problem Overview

River basin planning and management policies derived solely based on outcomes of physically-based watershed models usually fail to address severe social and environmental consequences (e.g., water resource conflicts). The crux of the problem is not the idea of using numerical models to assist policy making, but using the correct ones. The conventional physically-based models, on the one hand, lack the consideration of the uncertainty induced by heterogeneous human behaviors, which can lead to model overfitting while calibrating these models against observation data. On the other hand, these models lack the component to engage stakeholders and describe their interactions with the biophysical systems. Those models cannot take into account feedbacks from stakeholders who are affected by potential policies. Thus, failure to acknowledge the human behavioral uncertainty, and to describe interactions between human and biophysical systems can lead to poorly calibrated scientific models; the actual use of such model may end with unenforceable policies, or even worse, with unnecessary social conflicts.

In the new era of watershed management, in order to develop appropriate policies that can mitigate water conflicts and promote sustainable uses of water resources, policy makers not only need to understand the features of the physical systems, but also need to have good insights into the behaviors of humans interacting with the physical systems. An integrated modeling framework is proposed in an attempt to assist the design of sustainable watershed management policy by accounting for spatial and dynamic interactions between human behavior and biophysical processes, and modeling watersheds as coupled human and natural systems (CHNSs) (Pahl-Wostl, 2005; Hipel et al., 2007; Liu et al., 2007; An, 2012). The various actors in the human system can be modeled as a collection of autonomous decision-making and interactive entities, which are defined as agents. Agent-based modeling (ABM) or Multi-agent system (MAS) modeling has become popular methods to simulate human behaviors in various disciplines (Urban and Schmidt, 2001; differences between ABM and MAS are quite subtle and thus used interchangeably in the dissertation). This modeling framework allows modelers to focus on the attributes and behaviors of individuals, which otherwise may not be possible by using other modeling methodologies (Crooks and Heppenstall, 2012). Coupling a behavioral model with an environmental model can thus model a watershed as a CHNS by accounting for a wide range of relevant social, economic and environmental factors. Such coupled models can unveil hidden patterns in both human behavior and natural processes, capture emergent phenomena and thereby help us gain better understanding of the interactions between human and natural systems (Bonabeau, 2002; Monticino, et al. 2007; An, 2012). Thus, the main focus of this dissertation is on expanding the conventional watershed models with human behavioral simulation using the agent-based modeling framework.

However, modeling human behaviors is not simple, since the behaviors are not always rational due to limited information, cognitive abilities and time to make decisions (Simon, 1996). In this sense, the rationality of human behavior is bounded (see Kennedy, 2012). Thus, the main complexity of modeling human behaviors lies in the behavioral uncertainty arising from the bounded rationality. Two common approaches are usually used to derive the behavioral rules: 1) the rule-based approach and 2) the optimization-based approach. These approaches usually model human behavior explicitly with behavioral parameters. To address the uncertainty associated with these behavioral parameters, approaches like sensitivity analysis are usually used to investigate the impacts of the behavioral parameters on model outputs. These approaches can be very computationally expensive when the behavioral models become complex, not to mention that the behavioral models are usually coupled with physically-based environmental models for modeling watersheds as coupled human and natural systems. In addition, well-defined behavioral parameters, which are expected to capture the behavioral uncertainty, require modelers to have good domain knowledge. With the advancement in data science and computational infrastructure, an alternative approach is suggested deriving the agents' behavioral rules from a selective set of factors that are likely to affect agents' behavior and reflect bounded rationality of their behavior. However, there still exist some limitations in using a data-driven approach to derive behavioral rules primarily due to data availability and quality.

1.2 Research Objectives

The goal of the dissertation is to investigate the interactions between farmers and the groundwater system through developing and coupling a behavioral model with an environmental model using the High Plains aquifer hydrological observatory (HO) area as the study site. Specific objectives are developed as milestones along the way to reach the goal. These objectives are to

I. Apply an optimization-based approach to developing a behavioral model, and couple it with a groundwater model to simulate the High Plains aquifer within the HO area as a coupled human and natural system.

II. Propose a computational framework to improve the computational efficiency of the behavioral model and provide the network access to the coupled models with user scalability so as to support participatory modeling exercise and facilitate stakeholders' participation in groundwater resource management.

III. Develop another computational framework to deal with the computational intensity issue arising from global sensitivity analysis with the coupled behavioral model and groundwater model so as to capture the impacts of behavioral parameters on the outputs of the coupled models.

IV. Propose a new approach that combines data mining techniques (i.e., machine intelligence) with the expert domain knowledge (i.e., human intelligence) to derive the behavioral rules of agents as well as account for behavioral uncertainty.

Objective I aims to develop a behavioral model using an optimization-based approach by assuming that all agents are rational. Objective II attempts to address the system scalability and user scalability arising from the web-based application of the coupled behavioral model and environmental model. Quantification of the impacts of behavioral uncertainty of the coupled models is achieved through the computational framework developed by Objective III. Different from Objective I, Objective IV proposes a new approach to model human behavior and address the associated behavioral uncertainty from a different perspective. Achieving these objectives towards the goal of the dissertation not only involves the development of new methodological frameworks, but also presents the alternative means to model human behavior under the behavioral uncertainty.

1.3 Dissertation Outline

In Chapter II, a behavioral model is developed using optimization-based approach and a new computational framework is proposed to overcome the computational intensity of the coupled behavioral model and environmental model so as to support the coupled models as a service.

By following Chapter II, a cost-efficient approach is developed in Chapter III to measure behavioral uncertainty when the quantification of the impacts of human behavior becomes computationally intractable with the conventional approaches. By combining Chapter II and III, the goal to develop a behavioral model under the behavior uncertainty is achieved with the optimization-based approach.

Different from Chapter II and III that treat the behavioral uncertainty as the result of variations of behavioral parameters, Chapter IV attempts to propose a new approach for the design of the behavioral model under the condition of bounded nationality of human behavior.

Chapter V summaries the major findings and contributions of the present work, discuss the limitations as well as the future work.

References

An, L. (2012). Modeling human decisions in coupled human and natural systems: Review of agent-based models. *Ecological Modelling, 229*, 25-36. doi:10.1016/j.ecolmodel.2011.07.010

Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences of the United States of America, 99*, 7280-7287. doi:10.1073/pnas.082080899

Crooks, A. T., & Heppenstall, A. J. (2012). Introduction to agent-based modelling. *Agent-based models of geographical systems* (pp. 85-105) Springer.

Hipel, K. W., Jamshidi, M. M., Tien, J. M., & White,Chelsea C.,,III. (2007). The future of systems, man, and cybernetics: Application domains and research methods. *IEEE Transactions on Systems Man and Cybernetics Part C-Applications and Reviews, 37*(5), 726-743. doi:10.1109/TSMCC.2007.900671

Kennedy, W. G. (2012). Modelling human behaviour in agent-based models. *Agent-based models of geographical systems* (pp. 167-179) Springer.

Liu, J., Dietz, T., Carpenter, S. R., Alberti, M., Folke, C., Moran, E., . . . Taylor, W. W. (2007). Complexity of coupled human and natural systems. *Science, 317*(5844), 1513-1516. doi:10.1126/science.1144004

Monticino, M., Acevedo, M., Callicott, B., Cogdill, T., & Lindquist, C. (2007). Coupled human and natural systems: A multi-agent-based approach. *Environmental Modelling & Software, 22*(5), 656-663. doi:10.1016/j.envsoft.2005.12.017

Pahl-Wostl, C. (2005). Information, public empowerment, and the management of urban watersheds. *Environmental Modelling & Software, 20*(4), 457-467. doi:10.1016/j.envsoft.2004.02.005

Simon, H. A. (1996). *The sciences of the artificial* (Vol. 136). MIT press.

Urban, C., & Schmidt, B. (2001). PECS-Agent-Based Modelling of Human Behaviour. In *Emotional and Intelligent II-The Tangled Knot of Social Cognition, AAAI Fall Symposium*.

2 CHAPTER II

Design of a web-based application of the coupled behavioral model and environmental model for watershed management analysis using Hadoop

2.1 Introduction

Conventional river basin development planning and management fails to address severe social and environmental consequences (e.g., water resource conflicts), partially due to the lack of management policies and quantitative modeling tools available to engage stakeholders and describe their interactions with the biophysical systems (Barrow, 1998). A new integrated watershed management framework is proposed in an attempt to address these issues and promote watershed sustainability by accounting for the spatial and dynamic interactions between human and biophysical processes, and modelling watersheds as coupled human and natural systems (CHNSs) (Pahl-Wostl, 2005; Hipel et al., 2007; Liu et al., 2007; An, 2012). The various actors in the human systems can be modeled as a collection of autonomous decision-making and interactive entities, which are defined as agents. These agents follow certain behavioral rules in order to acquire new information, update their behaviors and adapt to the variations of the environment. They are also characterized by the rules to change these rules (North and Macal, 2007). Environmental models, on the other hand, are developed to simulate specific physically-based environmental processes in a natural system. Thus, coupling a multi-agent system (MAS) model with an environmental model can model a watershed as a CHNS by accounting for a wide range of relevant social, economic and environmental factors. Such coupled models can unveil hidden patterns in both human behavior and natural processes, capture emergent phenomena and thereby help us gain better understanding of the interactions between human and natural systems (Bonabeau, 2002; Monticino, et al. 2007; An, 2012).

Despite the recent progress made in modeling human behaviors for interactions with environmental systems, challenges still remain. For a modeler, one of the challenges lies in the computational intensity arising from running the MAS model with complex behavioral rules. In order to cope with this challenge, we usually resort to simplified behavioral rules for agents, and/or couple it with a lumped environmental model. In our prior work (Ng et al., 2011), a particular focus was to design an agent-based model (ABM) with relatively complex behavioral rules (utility optimization combined with learning), and couple it with a hydrologic-agronomic watershed model to simulate agents' decisions on crop and best management practice (BMP), as well as assess the environmental impacts of the decisions. One of the motivations for our current work is to address the computational intensity arising from running the coupled models with agents associated with complex behavioral rules. We propose to use the multithreaded programming to tackle the computational burden.

From the model application perspective, a large number of environmental models are developed but cannot be exploited by users other than the developers, which has affected the applicability of scientifically-based models for environmental research (Papajorgji et al., 2004). Using Web 2.0 technologies, design of a web-based application of environmental models becomes a viable trend for addressing the issues of data and model accessibility and service interoperability, and thereby increases the reusability of environmental models over computer networks (Papajorgji et al. 2004; Granell et al., 2010). For example, Castronova et al., (2013) demonstrate how to implement a hydrologic model, TOPography-based MODEL (TOPMODEL), as a web application using service-oriented architecture (SOA), which allows modelers to locate and couple different software components anywhere in computer networks. These individual components can then coordinate with each other via a certain protocol (such as HTTP protocol) to perform predefined tasks (Curbera et al., 2002; Huhns and Singh, 2005; Goodall et al., 2011). However, coupling the MAS model (e.g., with massive numbers of agents associated with evolving behaviors over time) with an environmental model (e.g., with long simulation time) by SOA is very challenging. The computational efficiency of the coupled models can be further affected when the large amount of data exchange used to describe the interactions between human and environmental systems becomes the bottleneck. In this chapter, we will discuss how to take advantage of the ease of implementation and flexibility of SOA while ensuring the computational efficiency of the coupled models as a web application.

In this chapter, a MAS model is developed based on our prior work (Ng et al., 2011), and coupled with a physically-based environmental model to simulate the watershed as a CHNS. We propose a framework to combine the multithreaded programming with Hadoop-based cloud computing for the parallel modeling of the MAS model as well as the scalable execution of the coupled MAS model and environmental model via the network. The remaining sections of the chapter are structured as follows. We begin with a brief introduction of MAS modeling and its applications for watershed management in the context of CHNS. Following that, we introduce parallelism, focusing on the computational patterns suitable for the MAS model and its parallel implementation. A Hadoop-based cloud computing scheme is then proposed to improve the user scalability of the web-based application. A real-world example that couples a MAS model for irrigation decision making with the Republican River Compact Administration (RRCA) groundwater model, is used to illustrate how to implement the coupled models in a Hadoop-based cloud computing environment (Illinois Cloud Computing Testbed, http://cloud.cs.illinois.edu/hardware.html) and deploy them as a web application. Finally, we conclude with some insights gained through the study.

2.2 Coupled MAS and Environmental Models

2.2.1 Multi-agent System (MAS) Model

The multi-agent system (MAS) modeling framework defines how to build numerical models, based on autonomous, interdependent and adaptive agents that follow prescribed behavioral rules. These models allow researchers to investigate the relationship between individual behaviors and collective social structures. Differing from conventional centralized approach, which assumes top-down control to handle decision-making processes, MAS is designed to follow a bottom-up, distributed approach (Becu et al., 2003; Reeves and Zellner, 2010; Ng et al., 2011), and thereby has the capability to assist in the spatial-temporal exchange of information in systems consisting of decentralized agents (An et al., 2005; Deissenberg et al., 2008; Robinson and Brown, 2009). As a result, MAS has gained popularity in both social and physical sciences over the last decade, moving from simple, theoretical exercises to more complex coupled social and biophysical models that describe a system, its components and the surrounding environments (Lansing, 1999; Reynolds et al., 2005; Ng et al., 2011). This dynamically coupled modeling approach can provide insight into progressive environmental feedbacks and subsequent societal responses (Reynolds et al., 2005).

For MAS modeling, the behavioral rules of agents define the ability of agents to learn and adapt to the environment. The number of agents and the associated behavioral rules usually determine the computational intensity of MAS. In the context of the coupled MAS model and environmental model, agents are usually defined as stakeholders with various behavioral rules operating on the shared common environmental resources (Feuillette et al., 2003; Monticino et al., 2007). The MAS model usually involves a large number of agents interacting with complex physically-based environmental models at multiple spatio-temporal scales (Epstein, 1999; Bennett and Tang, 2006; Tang et al., 2011). Their adaptations to the changing environment have to be simulated over a long period via the changes of behavioral rules (Gilbert and Troitzsch, 1999; Parker et al., 2003). As a result, the complexity of the MAS models and their intensive information exchange with environmental models can easily overwhelm the modeling efforts and force oversimplified representations of real-world dynamic phenomena (Haefner, 1992; Gong et al., 2013). This situation often leads to the simplification of the models or the reduction of the modeling scopes.

2.2.2 Computational Patterns and Parallelism

Rather than simplifying the coupled MAS model and environmental model by reducing the number of agents, defining them with simplified behavioral rules or coupling them with lumped environmental models, this study introduces parallel computing to ease the computational burden and solve the coupled MAS model and environmental model within a reasonable time. Parallel computing exploits the concurrency of the problem and solves the problem with multiple cores/processors. The problem can thereby be solved with less total wall-clock time than solving the problem on a single processor.

2.2.2.1 Agent and Repository Structural Pattern

A pattern language for parallel computing (Mattson et al., 2004; Massingill et al., 2005) is introduced to exploit the possibilities for parallelization in the coupled MAS model and environmental model. Specifically, the coupled MAS model and environmental model fit naturally to the agent and repository structural pattern defined by Our Pattern Language (OPL; Keutzer and Mattson, 2009). Different agents have their own tasks and communicate with each other on the basis of the common resources, including both information and natural resources. These common resources can be considered as repositories as shown by Figure 2-1. The hierarchical structure illustrates how parallelism can be introduced to the MAS model using the agent and repository pattern. At the upper level in Figure 2-1, different agents can have their independent tasks (e.g., crop choice and water use) associated with the common repository (e.g., shared resources) implemented in parallel, namely task parallelism. If we look into the individual behaviors of each agent (e.g., learning and optimization) at the lower level in Figure 2-1, we may be able to unveil the possibilities where other types of parallelism (e.g., data parallelism) can be introduced. For example, a large number of data samples can be generated in parallel and used to optimize the decision of each agent. As a result, depending on the tasks carried out by the agents, both task parallelism and data parallelism can be applied to improving the performance of the coupled MAS model and environmental model in terms of total wall-clock time.

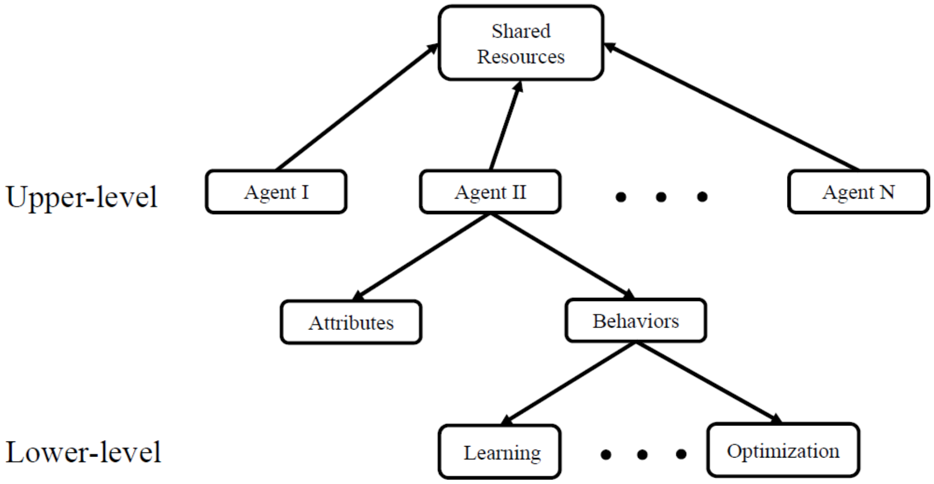


Figure 2-1 Architecture of the agent and repository pattern.

2.2.2.2 Parallel implementation

There exist a variety of ways to implement agent and repository structural pattern in parallel. For example, Tang et al. (2011) use message passing interface (MPI; Snir et al., 1998) to implement the parallel version of an ABM of land use opinions. Graphics Processing Units (GPU) programming is also used for the parallel modelling of ABMs (Tang and Bennett, 2012). However, these high-performance computing techniques are developed for specially designed CPU-GPU supercomputing resources. With recent advancements in multi-core technology for the x86/64 architecture, multithreaded programming which can initiate multiple threads to handle events concurrently on a single machine with less programming effort is becoming widely used. For example, a distributed platform for global-scale ABMs of disease transmission is built by Parker and Epstein (2011) using multithreaded programming. In this chapter, we will use the same programming technique for the parallel implementation of our MAS model.

In the context of the integration of MAS models with environmental systems, agents are usually associated with a large data set, including environmental data, social data and economic data, and execute independent tasks defined by their individual behavioral rules (Reeves and Zellner, 2010). Different tasks operate on the large shared data structure, which may consist of a set of text files, model solvers and databases. In order to improve the computational efficiency of the agents operating on the shared data structure, we try to execute the agents with independent tasks in parallel. However, unexpected results are likely to be generated due to simultaneous accesses to the shared data structure by multiple agents. Thus, the correctness of the program running in parallel must be guaranteed by making agent operations on the shared data structure thread-safe (Goetz et al., 2006). In other words, regardless of the scheduling of the threads generated by the runtime environment, a thread safe program running in parallel ought to generate the same results as the one running sequentially. To address the thread safety issue, the large shared data structure should be treated as a repository and a manager is needed to control the repository access by different agents and thus maintain the data consistency issue (Keutzer and Mattson, 2009). For the agent and repository pattern, one of the common approaches is to use relational database management systems (RDBMS) to manage the access to the repository, and ensure reliable database transactions. In addition, in the ideal ACID-compliant (A: atomicity; C: Consistency; I: Isolation; D: Durability) world, operations by different agents on the common repository managed by RDBMS can be programmed to be thread safe.

2.2.3 Model Coupling

As mentioned above, for the coupled MAS model and environmental model the computational intensity arises not only from the complexity of the coupled models themselves, but also from the data transfer between the MAS and environmental model. Data transfer usually occurs during the run-time to represent the interactions between human and natural systems so as to simulate the co-evolution of the two systems. Thus, they should be either tightly-coupled as a single model or treated as a single software component of the web application in order to avoid network latency (Goodall et al., 2011). The other components, such as web interface, visualization toolkits and web-based database can then be loosely coupled with the coupled models component using HTTP protocol (GET, PUT, POST, DELETE and HEAD) to reduce the complexity of the implementation of the model coupling and system debugging.

2.3 Cloud Computing with Hadoop

2.3.1 Cloud Computing

Most parallel computing paradigms for MASs or ABMs are evolved from the conventional supercomputing resources. For example, Da-Jun et al., (2005) design the parallel computing platform for the specific Beowulf-style clusters, and Tang et al., (2011) use supercomputer grids, TeraGrid, for the parallel agent-based modelling. However, in reality the time and monetary cost to access these computing resources can limit their applications. In addition, to acquire the technical knowledge and skills (such as those for dynamic load balancing, scheduling and synchronization) associated with the conventional high-performance computing can be challenging. In order to minimize the cost and circumvent these technical challenges, we propose a viable alternative to execute the coupled models in a cloud computing environment. Cloud computing is considered as a cost-saving means to bring unprecedented computing power which has huge potentials for water resources applications (Hunt et al., 2010; Liu et al., 2013). Its great advantage of on-demand access outweighs the supercomputer and makes it well suited for the web application in need of user scalability. For example, when the coupled MAS and environmental models are deployed as a web application, a variable number of users may expect to test the model simultaneously via the network without being affected by multiple accesses initiated by others. This can be satisfied if different model runs invoked by different users are executed independently on different machines, which in fact require easily scalable computing resources.

2.3.2 Apache Hadoop

Apache Hadoop is a commonly used open-source software framework used to create a cloud computing environment for large amounts of data storage and processing on clusters of commodity hardware (Apache Hadoop, 2009). This framework consists of two main components, Hadoop MapReduce (Dean and Ghemawat, 2008) and Hadoop Distributed File Systems (HDFS) (Borthakur, 2007). From a programmer’s point of view, MapReduce, similar to other Java libraries, is imported at the beginning of a program. One only needs to implement the map and reduce function defined by the "mapper" and “reducer” interface, and pass the input data into these functions. MapReduce takes over and ensures that the input data are distributed through the cluster, and computes the two functions across the entire cluster of machines. The details (i.e., parallelization, distribution of data and tolerance of machine failures) are hidden away from the programmer inside the library (Nielsen, 2009). We thereby do not need to be concerned about the load balancing and task rescheduling if any task fails. HDFS is used to store both the input and the output data from MapReduce jobs in the distributed file systems, ensuring data accessibility and consistency across the entire cluster. A detailed example is provided below to explain how we can leverage MapReduce for the parallel MAS modelling. As a result, we expect that the integration of the cloud computing with the Hadoop framework into the design of a web-based application of the coupled models will allow many users to execute the model simultaneously without an increase in latency.

2.4 Methodology

2.4.1 Case Study Site

The Republican River basin (RRB) in the U.S. Midwest is used as a case study. The Republican River originates in the high plains of northeastern Colorado, western Kansas and southern Nebraska. The basin covers approximately 25,018 square miles (~16 million acres) of the three states. The rapid development of irrigated agriculture in the basin in recent decades has resulted in a significant increase of groundwater use for irrigation and a declining water table. Due to the interaction between surface water and groundwater systems, the declining groundwater level has led to a trend of stream depletion that concerns policy-makers. Further, water conflicts arise because groundwater resources are shared by the three states. In order to understand groundwater use and the spatial impacts of groundwater pumping on streamflow, a comprehensive groundwater model, the Republican River Compact Administration (RRCA) groundwater model, was developed. The RRCA model uses MODFLOW-2000 (written in FORTRAN 77) with additional modules, and was developed and calibrated through the collaboration of the three affected states, the U.S. Geological Survey, and the U.S. Bureau of Reclamation (McKusick, 2003). The RRCA model is updated regularly, and data and code are freely available online (http://www.republicanrivercompact.org). In this study, a MAS model is developed to combine hydrologic, economic, and institutional factors into a detailed, cohesive and computationally tractable modeling framework for the simulation of farmers' irrigation behaviors on groundwater withdrawal. Thus, coupling it with the RRCA model can help us gain some insights into the interactions between famers' pumping behaviors and declining groundwater levels and stream depletion in the Republican River basin.

Using the case study site as an illustrating example, we will describe the design of the MAS model, the coupling of the MAS model with the RRCA model as a single software component, the procedure to execute the coupled MAS model and RRCA model in a parallel manner, and the integration with other components as a web-based application in the Hadoop-based cloud computing environment shown by Figure 2-2.

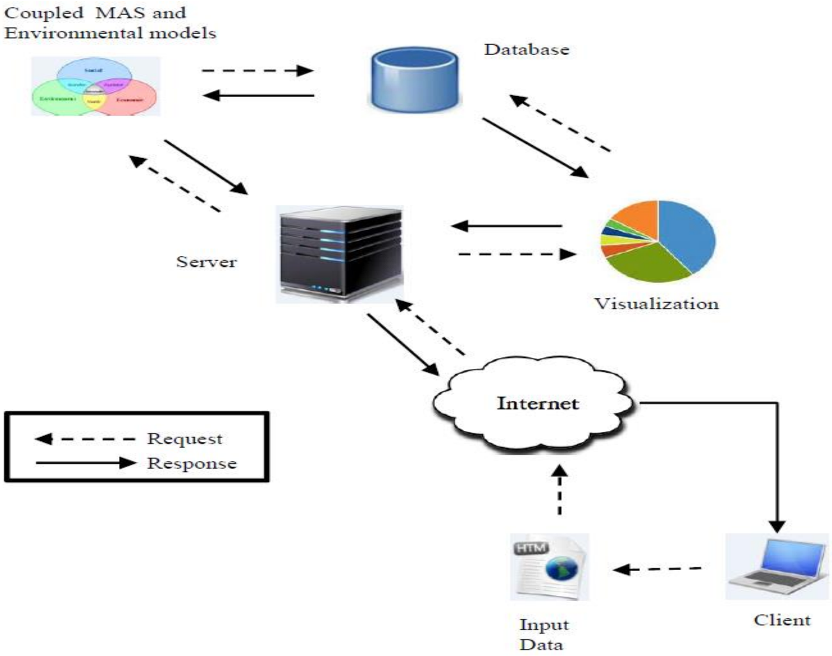
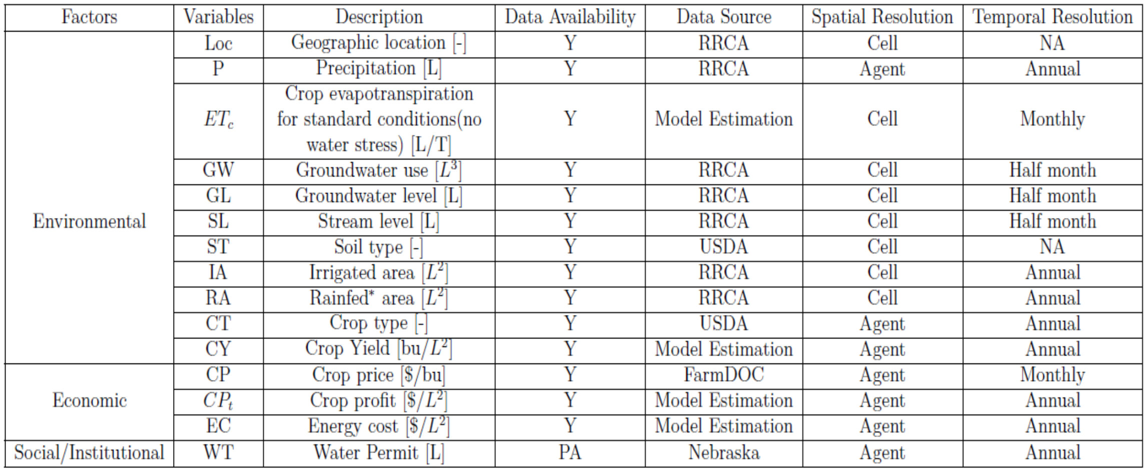


Figure 2-2 Architecture of the web-based application of the coupled MAS model and environmental model. Communications between the coupled models and the other components (database and visualization toolkit) are implemented via HTTP protocol.

Table 2-1 List of variables associated with agents' pumping behavior.



Factors	Variables	Description	Data Availability	Data Source	Spatial Resolution	Temporal Resolution
Environmental	Loc	Geographic location [-]	Y	RRCA	Cell	NA
	P	Precipitation [L]	Y	RRCA	Agent	Annual
	ET_c	Crop evapotranspiration for standard conditions(no water stress) [L/T]	Y	Model Estimation	Cell	Monthly
	GW	Groundwater use [L^3]	Y	RRCA	Cell	Half month
	GL	Groundwater level [L]	Y	RRCA	Cell	Half month
	SL	Stream level [L]	Y	RRCA	Cell	Half month
	ST	Soil type [-]	Y	USDA	Cell	NA
	IA	Irrigated area [L^2]	Y	RRCA	Cell	Annual
	RA	Rainfed* area [L^2]	Y	RRCA	Cell	Annual
	CT	Crop type [-]	Y	USDA	Agent	Annual
	CY	Crop Yield [bu/L^2]	Y	Model Estimation	Agent	Annual
Economic	CP	Crop price [$/bu]	Y	FarmDOC	Agent	Monthly
	CP_t	Crop profit [$/$L^2$]	Y	Model Estimation	Agent	Annual
	EC	Energy cost [$/$L^2$]	Y	Model Estimation	Agent	Annual
Social/Institutional	WT	Water Permit [L]	PA	Nebraska	Agent	Annual

Y: Yes; N: No; PA: Partially Available; NA: Not Applicable; RRCA: Republican River Compact Administration; USDA: U.S. Department of Agriculture; FarmDoc: Farm Decision Outreach Central; Nebraska: Department of Natural Resources, Nebraska; Cell: 1 mile by 1 mile. There are 13,220 cells in total which locate in 46 different agents. The number of the cells in the individual agent varies from few to approximately 1,000. Rainfed and Dryland are used interchangeably in this chapter.

2.4.2 Coupled MAS Model and RRCA Model

Pumping from the underlying aquifer is identified as the major water use practice for agricultural irrigation in the Republican River basin. Therefore, a better understanding of the impact of farmers' pumping behavior on water table and stream baseflow is of importance to resolve the water conflicts among the affected states. Here, the MAS model is used to simulate farmers' pumping decisions. The agent is defined as a county within the High Plains aquifer which underlies the Republican River basin. We assume that besides pumping from the same aquifer these agents have no explicit interactions with each other and their attributes are characterized by various factors from environmental, social and institutional perspectives, as shown by Table 2-1. Note that most of the data used by the MAS model are publicly accessible from the RRCA website (http://www.republicanrivercompact.org), U.S. Department of Agriculture (USDA), Farm Decision Outreach Central (FarmDOC) at University of Illinois and Department of Natural Resources, Nebraska, U.S. In addition, we estimate some factors at the desired spatio-temporal scale for the MAS model denoted by "Model Estimation" in Table 2-1. The data we used for the estimations are mainly obtained from Palazzo (2009) and Zhang et al. (2009 and 2010). We then develop a socioeconomic model to describe the behavioral rules of the agents by linking the environmental, institutional and economic factors with agents' decision on pumping (See Appendix A.1). We assume that agents behave strategically for the purpose of maximization of their individual utilities. In addition to this, agents' adaptation to the new environment is reflected in their ability to learn and update their behaviors over time.

2.4.2.1 Coupling the MAS model with the RRCA model

Coupling the multi-agent irrigation decision making system model with the RRCA groundwater model provides us with a way to investigate the water conflicts between the baseflow requirement for ecosystems and the demand for irrigation in the basin. Figure 2-3 (a) and (b) show spatial and temporal coupling processes between the MAS model and the RRCA groundwater model. For each agent, it communicates with hundreds of grids in the RRCA model within its boundary over the simulation period from year 1993 to 2006. Agent makes decisions on crop types, irrigated/rainfed areas and irrigation depth given the estimated crop prices and precipitation at the annual time scale. Then, the annual pumping rate for each agent is estimated and converted to the monthly pumping rates for each grid in proportion to their observed monthly pumping rates. Given the estimated pumping rates, the RRCA model updates the water table for each grid which is transferred to the MAS model, and then translated into crop profit function in terms of energy cost as shown by Figure 2-3 (c).

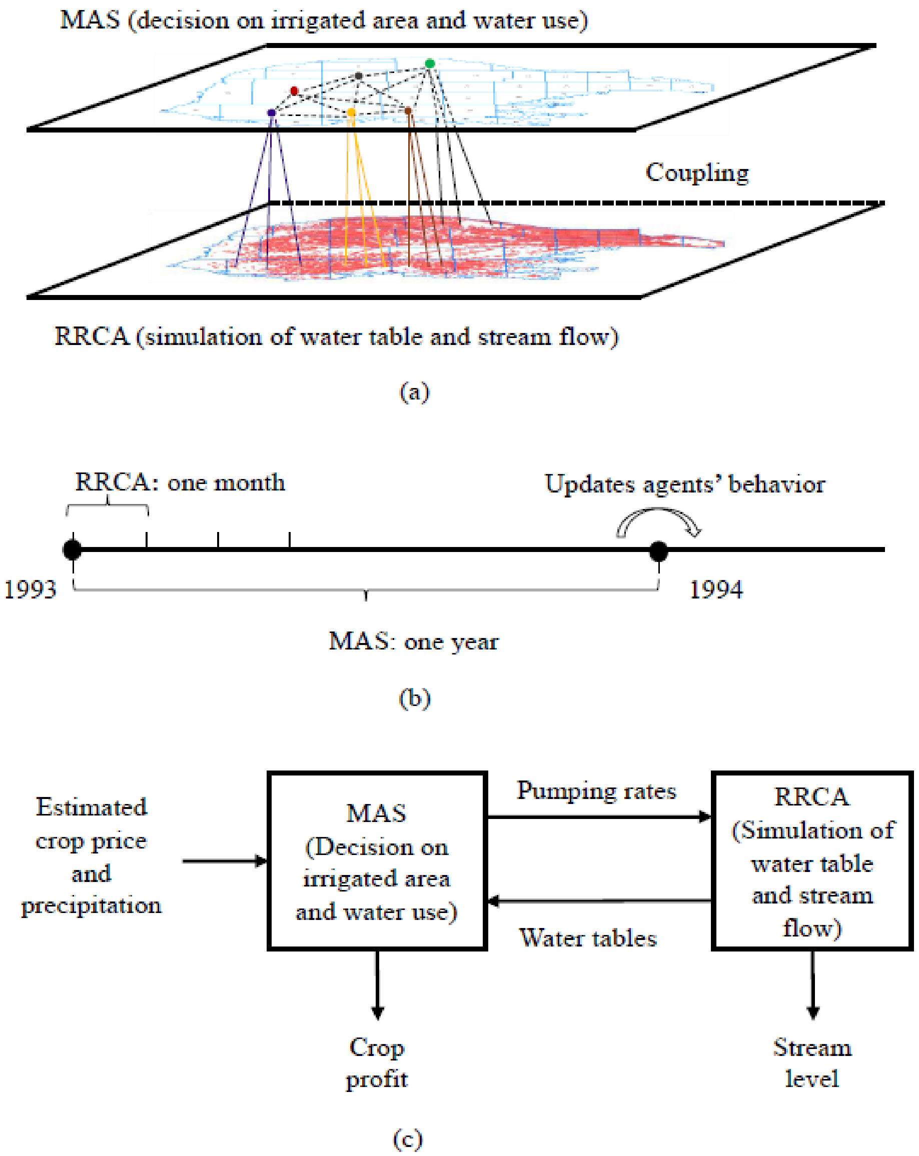

MAS (decision on irrigated area and water use)
Coupling
RRCA (simulation of water table and stream flow)
(a)
RRCA: one month
Updates agents' behavior
1993
1994
MAS: one year
(b)
Estimated crop price and precipitation
MAS (Decision on irrigated area and water use)
Pumping rates
Water tables
RRCA (Simulation of water table and stream flow)
Crop profit
Stream level
(c)

Figure 2-3 (a) Spatial coupling between the MAS model (upper part) and the RRCA groundwater model (lower part); (b) temporal coupling of the MAS model (with an annual time interval) and the RRCA groundwater model (with monthly time period); (c) exchange of pumping rates and water tables between the MAS model and the RRCA model.

2.4.2.2 Behavioral rules: utility maximization

Utility is defined as the agents' preference regarding crop profits in face of uncertainties. In our case, we assume agents deal with two kinds of uncertainties arising from both human and natural systems. The first one is the uncertainty of the future crop prices from the market, which we assume is one of the dominant factors associated with agents' decisions on crop types and planted acreage. The other uncertainty, coming from the physical environment, is precipitation, which is related to agents' decisions regarding irrigation depth. In our case, crop prices are determined only by the market and treated independently from precipitation. Agents update their perceptions about crop prices and precipitation via learning, and integrate them into their utility function.

Through this strategy, agents make optimal decisions on pumping in an attempt to balance the crop profits and risks associated with the uncertainties of crop prices and precipitation. In order to mimic this decision making process, we propose a Robust Optimization (RO) framework suggested by Mulvey et al. (1995) to define the utility function as follows:

$$U(\pi) = E(\pi) - \lambda \cdot Var(\pi). \tag{1}$$

Where E(π) and Var(π) are the expected value and variance of the crop profits, π, respectively. The agent's attitude towards the fluctuation of crop profits is denoted by λ. The higher the value of λ, the more risk averse the agent. Given the land and water availability constraints, we assume that all agents intend to maximize their utilities. Ng et al. (2011) used this form of utility to describe farmer agents who decide crop choices. In our case, the utility maximization problem is solved using a two-stage stochastic and deterministic optimization algorithm, which helps agents determine the crop types, optimal irrigated/rainfed area and irrigation depth under the uncertainty of crop prices and precipitation at the annual time scale (see Appendix A.1).

2.4.2.3 Bayesian learning

Developing a learning process is essential for agents to adapt to the ever-changing external environment. In our case, it is assumed that individual agents predict the crop prices and precipitation for the crop growing season before planting the crop. They have their own perceptions about crop prices and precipitation derived from their past experiences, namely prior knowledge. In addition to their prior knowledge, they also observe the crop prices and precipitation right before the crop planting season. In our study, we incorporate an agent's prior knowledge and their observations to derive their posterior knowledge of crop prices and precipitation, which can then lead to adaptive decisions over time.

Bayesian statistics is used to simulate the learning process. Bayes' theorem specifies how prior beliefs (in our case, the prior knowledge of crop prices and precipitation) should be combined with the diagnosticity of the evidence (agents' current observations), denoted by the likelihood function (Kahneman, 2011, p154). The simulation of the learning process is an extension of the work by Ng et al. (2011). For both crop prices and precipitation, we assume that their likelihood function follows the normal distribution:

$$p(D \mid \mu, \sigma^2) = \frac{1}{(2\pi)^{n/2}} (\sigma^2)^{-n/2} \exp(-\frac{1}{2\sigma^2}[n\sum_{i=1}^{n}(x_i - \bar{x}) + n(\bar{x} - \mu)^2]). \qquad (2)$$

Where $D = (x_1, \cdots, x_i, \cdots, x_n)$ is the observed data, which are an independent and identically distributed (IID) sequence and $\bar{x}$ is the mean of the sequence. μ and σ^2 are the mean and variance of the likelihood function. A suitable conjugate prior, normal-inverse-chi-squared ($NI\chi^2$) prior as the product of normal distribution (N) and inverse-chi-squared distribution (χ^{-2}) is used (Murphy, 2007):

$$p(\mu, \sigma^2) = NI\chi^2(\mu_0, \kappa_0, \nu_0, \sigma_0^2) = N(\mu \mid \mu_0, \sigma^2 / \kappa_0) \cdot \chi^{-2}(\sigma^2 \mid \nu_0, \sigma_0^2). \qquad (3)$$

Where μ_0 is the prior mean and κ_0 is how strongly we believe the prior mean; σ_0^2 is the prior variance and ν_0 is how strongly we believe this. The hyperparameters μ_0 and σ^2 / κ_0 can be interpreted as the location and scale of μ, and the hyperparameters ν_0 and σ_0^2 as the degrees of freedom and the scale of σ^2. We can obtain the posterior distribution of prices and precipitation via Bayes' theorem (Lee, 2004, P67):

$$\begin{aligned}
&p(\mu, \sigma^2 \mid D) = NI\chi^2(\mu_n, \kappa_n, \nu_n, \sigma_n^2) \propto p(\mu, \sigma^2) p(D \mid \mu, \sigma^2) \\
&\qquad \mu_n = \frac{\kappa_0 \mu_0 + n\bar{x}}{\kappa_n} \\
&\qquad \kappa_n = \kappa_0 + n \\
&\qquad \nu_n = \nu_0 + n \\
&\qquad \sigma_n^2 = \frac{1}{\nu_n}(\nu_0 \sigma_0^2 + \sum (\mathrm{x}_i - \bar{x})^2 + \frac{n\kappa_n}{n + \kappa_n}(\mu_0 - \bar{x})^2).
\end{aligned} \qquad (4)$$

Where μ_n is the posterior mean and κ_n represents the level of confidence in the posterior mean; σ_n^2 is the posterior variance and ν_n reflects the level of confidence in the posterior variance. By the Bayesian approach, the model allows agents to adjust their annual predictions of the expected crop prices and precipitation as they make new observations, which will further impact agents' decisions on groundwater pumping for irrigation.

2.4.3 Software Design of the Coupled Models

The coupled MAS model and RRCA groundwater model are implemented in the object-oriented language, Java. The implementation and interaction of different Classes are described by the unified modeling language (UML) (Muller, 1997) as shown by Figure 2-4. The coupled models are initialized by the two Java Classes, *"RRBProjectViewer"* and *"RRBProjectController"*. Each agent's learning and optimization process is then controlled by the Java Class, "*ExecuteAgent*". This class acts as the brain of the agent, invoking the *getPosterior()* method in the Java Class, "*BayesianLearning"* to update the agent's prediction of the future crop prices and precipitation on the basis of their prior knowledge about them and the current observations. Based on their newly updated estimations of crop prices and precipitation, the *getPlantedAreaMethod()* and *getAgentWaterUseandCropArea()* method defined in the Java Classes, "*AgentPlantedArea*" and "*AgentWaterUseArea*" are invoked by the "*ExecuteAgent*" to solve the two-stage stochastic optimization problem (see Appendix A.1), and help agents to determine the crop types, optimal irrigation/rainfed area and groundwater usage. An open source software package with application program interfaces (APIs) written in Java, called the Java Native Interface (JNI, https://projects.coin-or.org/Ipopt/wiki/JavaInterface) for Interior Point OPTimizer (Ipopt) (Wächter and Biegler, 2006) is tightly integrated into the MAS model to solve our large-scale nonlinear optimization problems (NLPs).

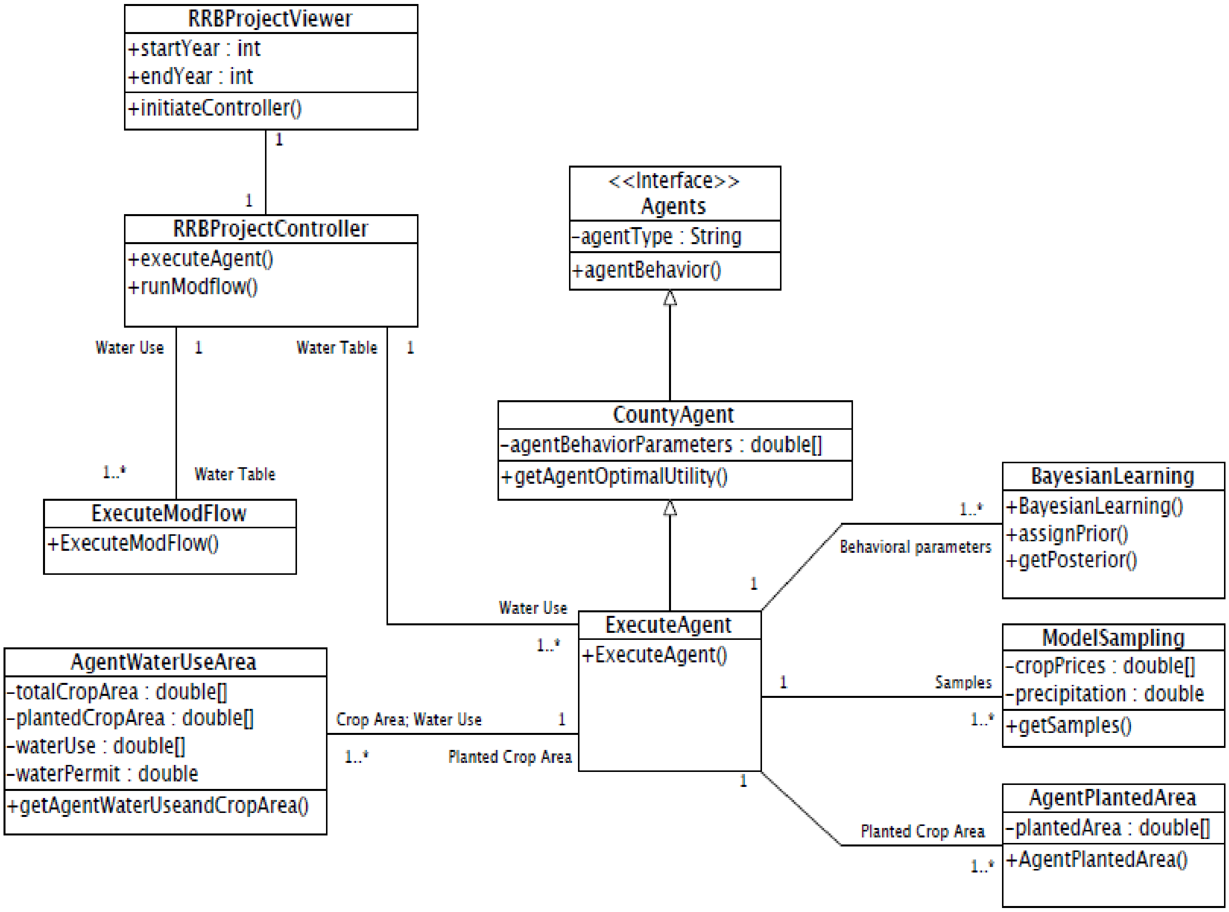


Figure 2-4 UML description of the coupled MAS model and RRCA model. The coupled models consist of one Interface and eight Classes.

Following the optimization, agents then determine the annual pumping rates. Their decisions on pumping rates are the key driving force to the RRCA model, which is invoked by the Java Class, "*ExecuteModFlow*"through calling the MODFLOW executable file. The outputs of the RRCA model are then used as the feedback to the MAS model. One of the key feedbacks is the water table, which can be used to evaluate the impacts of agents' pumping decisions on groundwater. The water table is then translated into the crop revenue function in terms of energy cost (See Appendix A.1), and thereby affects agents' decisions on pumping in the consecutive year. The communication between MAS and RRCA model is implemented through data exchange by input/ output files to/from each other.

In addition to the computational complexity, running the coupled MAS model and RRCA model also generates a huge variety of data for each agent, including environmental and socioeconomic data listed in Table 2-1. In order to store and manage the data for the agents, we decided to use the popular open-source relational database, MySQL. Given that the MAS model is coded in Java, Java Database Connectivity (JDBC) APIs (Poo et al., 2007) are used by the Java Classes, "*BayesianLearning"* and "*ExecuteAgent*" to store, query and update data in the MySQL database, which ensures that all transactions occur in an atomic fashion. Thread safety is thereby guaranteed when multiple agents want to access and operate on this common repository.

2.4.4 Parallel Version of the Coupled Models

2.4.4.1 Profiling and Performance Analysis

Coupling the MAS model defined by complex behavioral rules with the large-scale physically-based groundwater model results in a high computational intensity. Originally it took nearly four hours to run the coupled models sequentially over the fourteen year simulation period. We identify two main performance bottlenecks according to the profiling and performance analysis: 1) 46 agents execute learning and optimization in sequence, and each optimization needs to call the external MATLAB nonlinear optimization solver, *fmincon* function; 2) For each agent, a large amount of samples are sequentially drawn from the posterior distribution of the crop prices and precipitation to calculate the sample mean and variance for solving the stochastic optimization problem by the *getSamples()* method in the Class, *"ModelSampling"*. Notice that the second bottleneck is embedded in the learning process. Currently 74% of the original sequential runtime is related to these two bottlenecks.

2.4.4.2 Parallel Implementation

According to the pattern languages for parallel computing (Mattson et al., 2004) and the features of our MAS model, we decide to parallelize the MAS model with task parallelism for the first bottleneck, and apply both task and data parallelism for the second bottleneck as follows. For each agent, their tasks of variable initialization, optimization and learning are independent from other agents. We thereby create workers with the Java concurrent package to execute the tasks for each individual agent. The outputs from each agent are collected as the inputs to the groundwater RRCA model after the execution is completed. The second bottleneck comes from the use of the Monte Carlo method to generate samples from the posterior distribution of crop prices and precipitation. Note the sequential invocation of the random number generator and the sequential execution of the corresponding function evaluation with these random numbers cause the slowdown of the program. Given the assumption that the samples are independent and identically distributed, we then introduce task parallelism to sample generation, and data parallelism to the evaluation functions to improve the speedup.

In this study, due to data limitation, the interactions between agents are not implemented explicitly, but implicitly through the sharing of the groundwater via the RRCA model. It is assumed that each agent optimizes utility individually; however, the connection among agents' decisions on pumping for irrigation is captured by the conditions of water available for pumping and pumping cost (depending on water table), which is simulated by the RRCA model coupled with the MAS model. It should be admitted that this correspondingly reduces the computational complexity of the parallel implementation of the MAS model with multithreaded programming. If the communication between agents is explicit the multithreaded programming will be complicated, since individual threads need to handle the information shared by the agents. In the literature there are some studies that handle the communications between agents for the parallel implementation of a MAS model. For example, Parker and Epstein (2011) applied the priority queues (TreeMaps) to tackle the scheduling of simulation events; Tang et al. (2011) used the ghost zones approach to address the exchange of information between agents.

Thread safety is extremely crucial to the correctness of a parallel program. We use two approaches to address code thread safety. First, the variables associated with individual agents, such as agent's behavioral parameters, were originally defined as Class variables, which are visible to other agents. Now, we redefine them as local variables, becoming invisible to other agents. Second, for the variables shared by different agents such as some socioeconomic and environmental factors listed in Table 2-1, we categorize variables into three groups: 1) immutable objects; 2) read-only objects; 3) modifiable objects. Only the third type, modifiable objects, has thread safety concerns. In addition to the use of RDBMS, we also use two other methods to ensure the thread safety. The first method is to replace some non-atomic data structures with the built-in concurrent collections. For example, the data structure, "*MultiKeyMap*" in Java can be replaced by the atomic data structure, "*ConcurrentHashMap*". The second method uses synchronized methods to restrict the shared data access among threads and thereby avoid accessing the same resources by the multiple threads, so-called race conditions. In this method, we try to improve the execution time by reducing the lock duration only to the part of the Java functions that can cause race conditions, rather than synchronizing the entire functions in the Class.

2.4.5 Coupled Models as a Web-based Application

When the coupled models are deployed as a web-based application, it is expected that multiple users can run the model over a network simultaneously without incurring increased latency caused by each other, namely user scalability. We thereby propose to execute the individual instances of the coupled models in parallel on a multi-core node, and different model runs invoked by different users over the web interface are distributed and executed across different nodes using the Apache Hadoop software framework. Various input files required by the agents are distributed locally to each map task from the HDFS. Each agent is executed with multiple threads on a single node within the map phase. Different instances associated with different scenario IDs are executed on other available nodes of the cluster. Figure 2-5 shows model inputs stored in the HDFS are copied to the local system when the map phase is invoked. The coupled models are then executed with multiple threads over fourteen years and no execution occurs in the reduce phase. Another instance can be running on another node in the same manner.

The coupled models use some shared libraries not available on the Illinois Cloud Computing Testbed. We thereby distributed the libraries through the distributed cache (Hadoop Map/Reduce tutorial, 2013). All shared libraries are placed in a zip file and added to the distributed cache. The paths to the shared libraries are added as Hadoop system environmental variables, which make the libraries available to the programs written in Java as shown by Figure 2-5. Only a few selected outputs from the MAS model are then saved into the web database and the rest are save in files in HDFS.

Adding the web interface and visualization toolkits is about making it easier to start and manage the coupled models running in a Hadoop-based cloud environment, and then to visualize the model results as shown by Figure 2-6. Users can assign different values to the behavioral parameters ($k_{pr}, v_{pr}, k_{prep}, v_{prep}$ and λ) for different agents (e.g., in different states located in the study basin) via the web interface on the client side. These behavioral parameters are associated with agents' learning and utility maximization as mentioned above. The input information is stored into the database with a unique scenario ID via HTTP POST function. JavaScript Object Notation (JSON) is used as data exchange format between the database on the server side and the web interface on the client side. After the execution is completed, users can view the results, including crop-related information and the water table of selected points along the Republican River. Asynchronous JavaScript and XML (AJAX) is applied to allowing the clients to query the information from the database via the web interface, and Google Charts APIs (Google Charts, https://developers.google.com/chart/) are used as visualization toolkits to visualize data on the web.

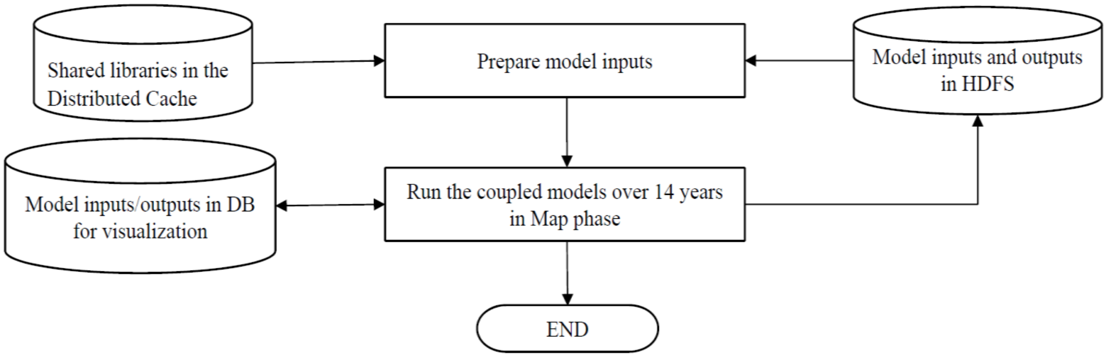

Shared libraries in the Distributed Cache
Prepare model inputs
Model inputs and outputs in HDFS
Model inputs/outputs in DB for visualization
Run the coupled models over 14 years in Map phase
END

Figure 2-5 Overview of the execution of the coupled MAS model and RRCA model with the Apache Hadoop framework. The shared libraries are distributed through the distributed cache.

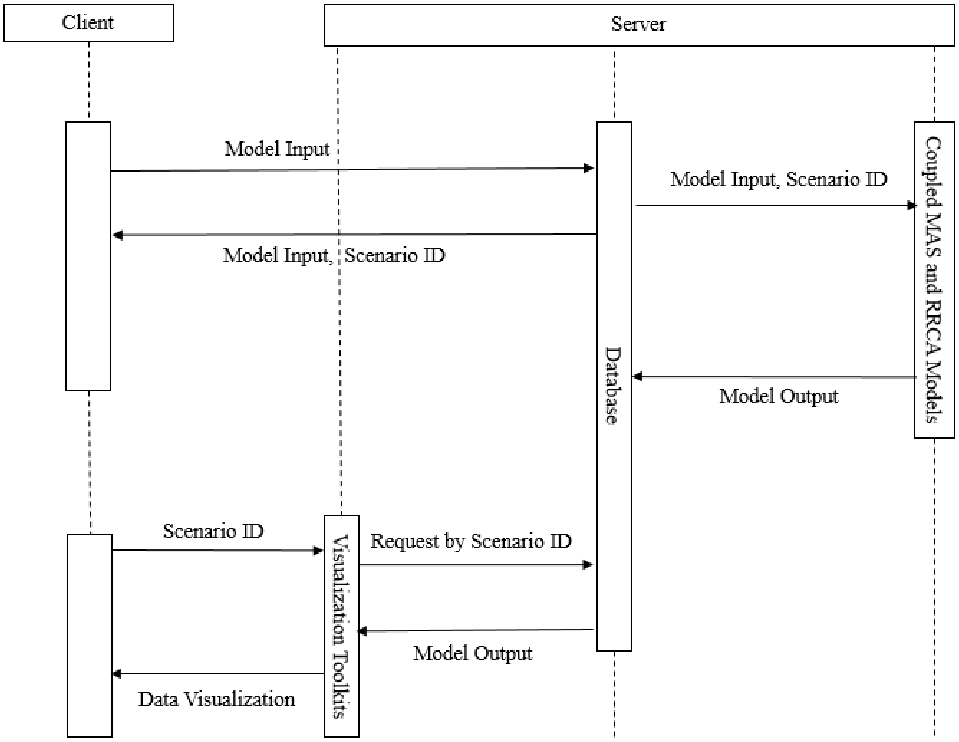


Figure 2-6 The client-server model: data exchange between the web interface on the client side and the Database, coupled MAS and RRCA models and visualization toolkits on the server side.

2.5 Results and Discussions

One challenging issue to make the coupled MAS model and RRCA model useful for scientific research is to improve its computational efficiency. Initially, it took nearly four hours to execute the coupled models over a fourteen year study period. We found that one of the major culprits of the computational burden was the optimization solver. Our software originally used the MATLAB *fmincon* function to numerically solve the two-stage stochastic and deterministic optimization (see Appendix A.1). This method needs to export the *fmincon* function as a Java .jar file via MATLAB Builder™, JA (MATLAB Builder™, JA, http://www.mathworks.com/products/javabuilder/), and integrate it as a Java Class into the coupled models. However, the overhead to initiate the external optimization solver from the coupled models is too large and there are potential concerns regarding licensing when using a commercial solver for the application. Because of these concerns, we incorporated the open-source optimization solver, Ipopt. We test both solvers using the same inputs drawn from agents, showing the results from Ipopt are the same as the MATLAB solver, but with Ipopt the total running time of the coupled models has been reduced from nearly four hours to one hour for 46 agents over fourteen years (i.e., the runtime for each individual agent with the complex behavioral rules is about one second with Ipopt). In addition, the MAS model is also very stable with Ipopt, resulting in nearly no runtime errors that prevent the model from working properly.

System scalability is an important attribute associated with multi-processor systems. It is used to measure the effects of multithreaded programming in comparison with the sequential implementation of the program. We use speedup to define system scalability and measure the program’s running time with different numbers of threads from one to eight on an eight-core workstation (4 Intel Xeon(R) E7329 @ 2.13GHz Dual Core Processors, 32GB DDR3 memory and network file system). Two measurements have been carried out to measure the speedup of the parallel version of the program against the original sequential version. The first measurement is used to measure the speedup of the section where we implement task and data parallelism. The second is used to measure the overall speedup. For each statistic, we measure the execution time five times and then average them. The speedup is then calculated by Amdahl's law (Amdahl, 1967). Figure 2-7 (a) shows a nearly linear scalability with the increase of the number of cores for the first measurement. Under the assumption that there exist no explicit interactions between the agents, the speedup should be linear to the number of cores when the overhead to initiate a single thread is not taken into account. But, in fact, the speedup will level off when the overhead offsets the runtime reduced by the additional thread. Figure 2-7 (b) shows the overall speedup for the coupled models. Notice that there exists an increasing gap between the overall speedup and the linear line in Figure 2-7 (b). This observation can be explained by that the running time for the parallel portion decreases significantly (from 91.4% of total execution time for a single core to 57.5% of the total execution time for eight cores) while the runtime for the remaining portions of the program (e.g., the RRCA model) remains the same. The total running time of the coupled models is thereby reduced from one hour to twelve minutes on an eight-core desktop machine.

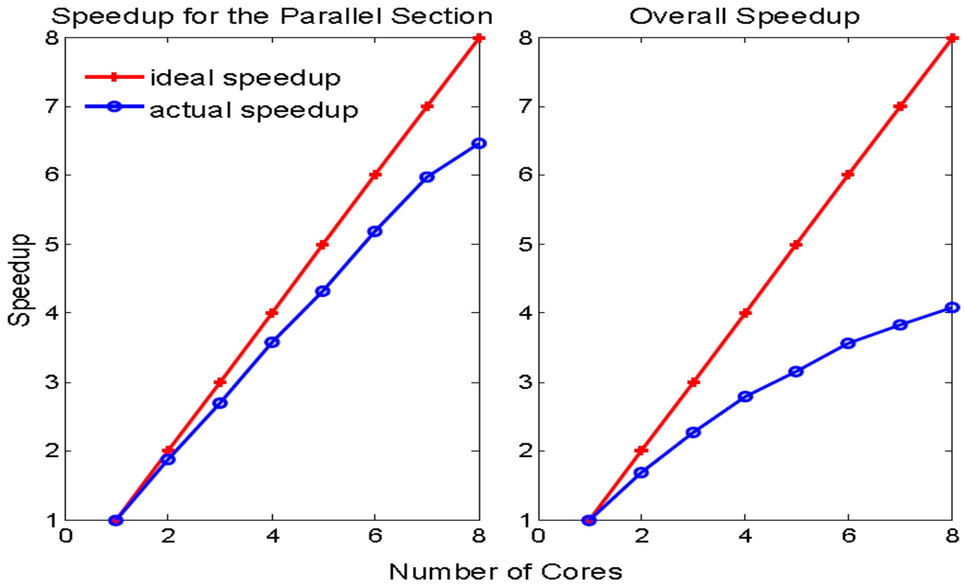
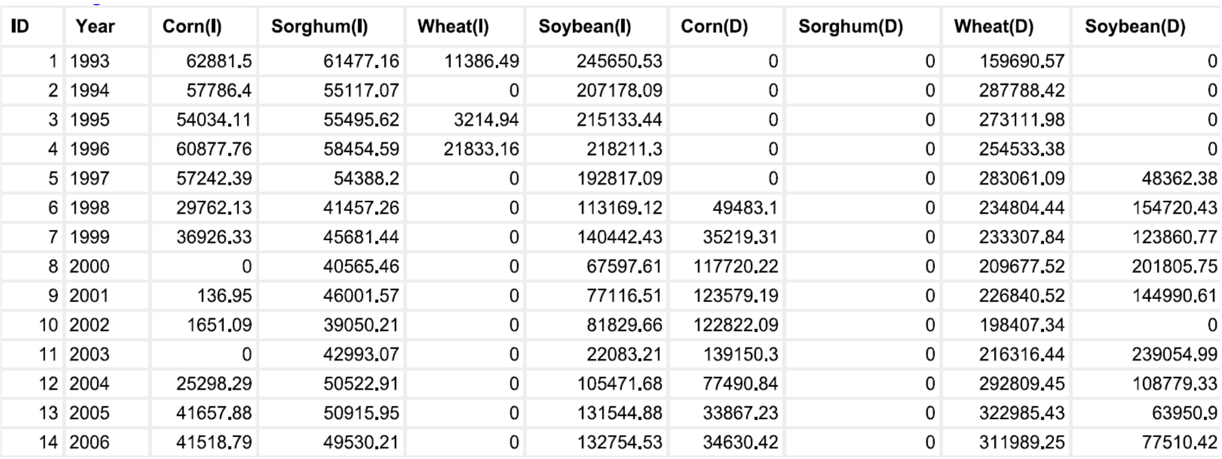


Figure 2-7 (a) Speedup for the parallel section (left) and (b) Overall speedup (right).

For the sake of user scalability, we decided to deploy the web-based application of the coupled models in a Hadoop-based cloud computing environment. The MAS model lends itself nicely to the MapReduce framework under the assumption of no direct communications between agents in the MAS model. We originally designed two approaches to execute the coupled models with the Apache Hadoop framework. For approach one, in the map phase, agents in the MAS model execute their tasks independently in different nodes, and outputs from the agents are combined in the reduce phase as the inputs for the RRCA model. For approach two, rather than spreading the agents over different nodes, as mentioned above, we execute the simulation of all agents over fourteen years in parallel on the same node in map phase. Likewise, the outputs are used as inputs for the RRCA model executed in the map phase as well. As a result, no execution occurs in the reduce phase. Consider that the first approach has a significant amount of overhead associated with starting a MapReduce job, and agents also require very large amount of input data that must be distributed locally to each map task. We therefore believe that the second approach with multiple threads is more efficient by taking advantage of data locality while executing the MAS model, but more analysis needs to be done to confirm our hypothesis.

The coupled MAS model and RRCA model are deployed as a web application for better accessibility. The web interface (http://waterproject.web.engr.illinois.edu/) can help clients access the coupled models running on the cluster, in addition to viewing and comparing model results. For example, Figure 2-8 shows the crop information for a specific agent under the scenario where

$k_{pr} = 3, v_{pr} = 5$, $k_{prep} = 4$, $v_{pr} = 5$ and $\lambda = 10$ for all the agents in Colorado. The agents in Kansas and Nebraska use the default values ($k_{pr} = 5, v_{pr} = 40$, $k_{prep} = 5$, $v_{pr} = 40$ and $\lambda = 2.5$). To be more specific, Figure 2-8 (a) shows the distribution of total/irrigated crop area over the simulation period from year 1993 to 2006. For readability, the results in Figure 2-8 (a) are represented as a stepped-area graph as shown by Figure 2-8 (b). Users can also select a specific year and view the water use, distribution of crop area and crop profit of that year as shown by Figure 2-8 (c). In addition, users can compare the impact of pumping behavior on crops and the water table with different settings of behavioral parameters under different scenarios. Figure 2-9 shows the distribution of crop area under the specific scenario ($k_{pr} = 5, v_{pr} = 5$, $k_{prep} = 5$, $v_{pr} = 5$ and $\lambda = 5$ for Nebraska and the other agents use the default values). In comparison with the results in Figure 2-8 (b), wheat is the dominant crop planted in the dryland area in Figure 2-9.

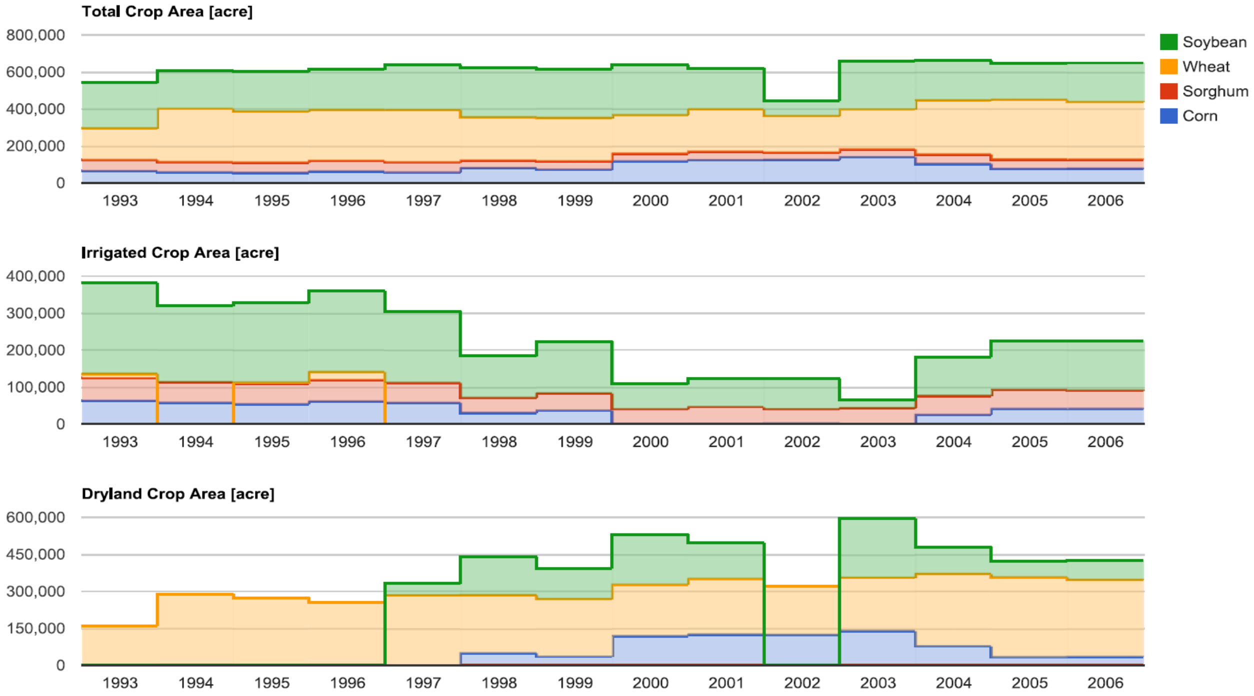


ID	Year	Corn(I)	Sorghum(I)	Wheat(I)	Soybean(I)	Corn(D)	Sorghum(D)	Wheat(D)	Soybean(D)
1	1993	62881.5	61477.16	11386.49	245650.53	0	0	159690.57	0
2	1994	57786.4	55117.07	0	207178.09	0	0	287788.42	0
3	1995	54034.11	55495.62	3214.94	215133.44	0	0	273111.98	0
4	1996	60877.76	58454.59	21833.16	218211.3	0	0	254533.38	0
5	1997	57242.39	54388.2	0	192817.09	0	0	283061.09	48362.38
6	1998	29762.13	41457.26	0	113169.12	49483.1	0	234804.44	154720.43
7	1999	36926.33	45681.44	0	140442.43	35219.31	0	233307.84	123860.77
8	2000	0	40565.46	0	67597.61	117720.22	0	209677.52	201805.75
9	2001	136.95	46001.57	0	77116.51	123579.19	0	226840.52	144990.61
10	2002	1651.09	39050.21	0	81829.66	122822.09	0	198407.34	0
11	2003	0	42993.07	0	22083.21	139150.3	0	216316.44	239054.99
12	2004	25298.29	50522.91	0	105471.68	77490.84	0	292809.45	108779.33
13	2005	41657.88	50915.95	0	131544.88	33867.23	0	322985.43	63950.9
14	2006	41518.79	49530.21	0	132754.53	34630.42	0	311989.25	77510.42

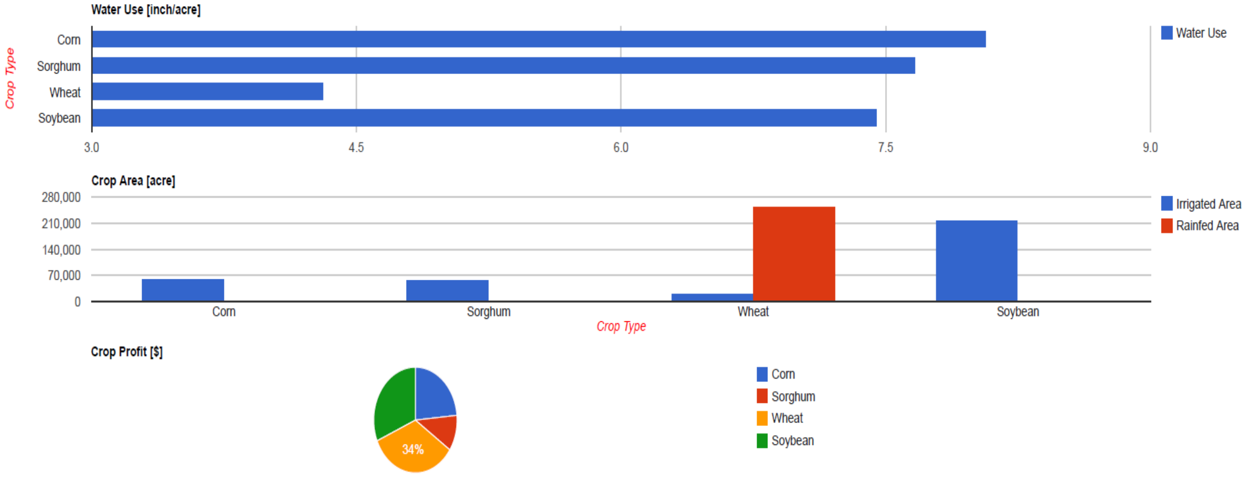
(a)

Figure 2-8 (a) Irrigated (I) and dryland (D) crop area [acre] for agent 18 over the simulation period from year 1993 to 2006; (b) distribution of total, irrigated and dryland crop area [acre] for different crops for agent 18 displayed in a stepped area graph; (c) water use [inch/acre], crop area [acre] and the percentage of crop profit [$] for different crops for agent 18 in year 1996.

Figure 2-8 (continued). (Caption shown on previous page.)



(b)



(c)

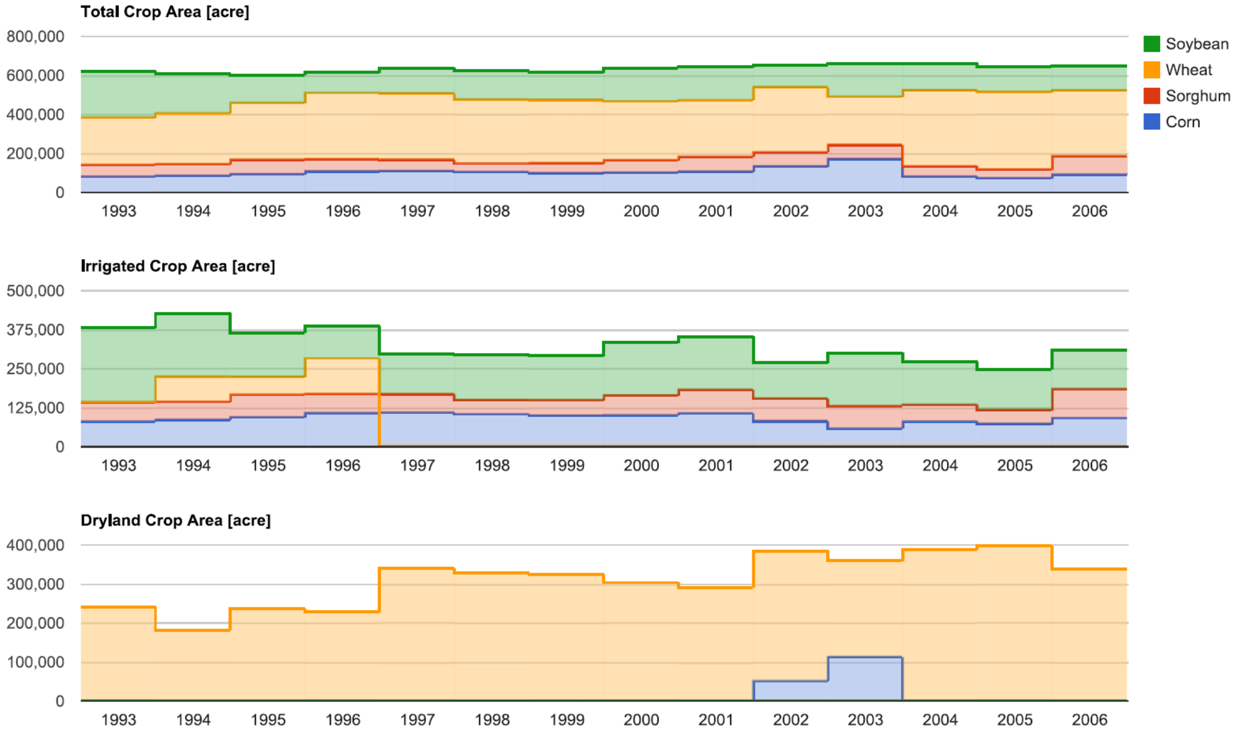


Figure 2-9 Distribution of total, irrigated and dryland crop area [acre] for different crops for agent 18 for the specific scenario ($k_{pr} = 5$, $v_{pr} = 5$, $k_{prep} = 5$, $v_{pr} = 5$ and $\lambda = 5$ for the agents in Nebraska and other agents use the default values).

Moreover, multiple users can invoke instances of the coupled models via the web interface and these instances will be distributed and executed on different nodes through the MapReduce framework when nodes are available. However, one should be aware that during execution these instances access the shared web-database simultaneously where the provider sets a hard constraint on the number of database connections. As a result, initiating many instances simultaneously can easily exceed the constraint and lead to the execution failure of each individual instance.

2.6 Summary and Conclusions

We have developed a web-based application of a coupled MAS model and environmental model for a case study of irrigation and its environmental impacts in the Republican River basin. A multi-agent system model is designed to simulate the agents' pumping behaviors, and it is coupled with the physically-based RRCA groundwater model. The MAS model, which incorporates self-learning and utility maximization for the various agents, simulates agents' decisions on the crop types, optimal irrigated/dryland area, and irrigation depth at the annual time scale; meanwhile the RRCA model simulates groundwater state (e.g., water table). As a result, the coupled models allow us to investigate the interactions between the agents' pumping decision-making and the groundwater system.

This chapter focuses on the computational complexity of the coupled models, which could limit the usefulness of the models. To tackle the computational issue, we first integrate the open source optimization solver, Ipopt, into MAS. The test result shows that the execution time of the coupled models running in sequence is reduced from nearly four hours with MATLAB solver to one hour with the Ipopt solver. We then introduce multithreaded programming to ease the computation intensity of the coupled models by executing the simulation of the agents in parallel using the agent and repository pattern from OPL. A database is used to store and manage data to ensure the thread safe access to the shared data structures. Other approaches such as concurrent data structure and synchronized methods are used to avoid race conditions and thus ensure the thread safety while executing the MAS model in parallel. The speedup test in Figure 2-7 shows that the execution time of the parallelizable part of the program is reduced from 91.4% of the total execution time for a single core to 57.5% for eight cores. The total running time of the coupled models is reduced by 80%, from one hour down to twelve minutes on an eight-core node. The significant improvement of the computational efficiency for the coupled models opens up possibilities for implementing complex models for real world applications. In particular, it empowers users to conduct even more time-consuming tasks with complex models, such as sensitivity analysis and model calibration.

A web application will allow many users to access the application over a network. Cloud computing, using the Apache Hadoop software framework, including MapReduce and HDFS, brings us on-demand access to the unprecedented computational power via the network. As a result, the Hadoop framework integrated in the design of the web application enables the coupled models to handle a large number of invocations through the web interface. As a demonstration, the coupled models are deployed to the Illinois Cloud Computing Testbed and different instances initiated by different users via a web interface should be primarily distributed and executed on different nodes by the MapReduce framework when nodes are available. It is found that the Hadoop-based cloud computing scheme allows the coupled models to effectively respond to a number of requests by the users. Moreover, the web application is facilitated by loosely coupling the model with other components via HTTP protocol, including web interface, web-based database and visualization toolkit.

In summary, our experimental work takes advantage of advancements in computer technologies (multithreaded programming, Hadoop-based cloud computing and web 2.0) to build a platform that provides network access to a physically realistic and computationally efficient coupled MAS model and environmental model. This chapter presents an initial effort of modeling CHNSs as a web application which facilitate an online dissemination of the models and the results, and then support participatory modeling exercises. Although the work presented in this chapter is tailored to the Republican River basin case study, the design of the coupled models can be used as a reference for other cases that involve the interactions between human and natural systems. The framework which combines multithreaded programming with Hadoop-based cloud computing and managed from a web application is generally applicable to models which require system and user scalability. Multithreaded programming can be applied to improving the computational efficiency of the single instance of the model (i.e. system scalability). Hadoop-based cloud computing provides on-demand computational power to execute multiple instances of the model simultaneously (i.e. user scalability).

References

Apache Hadoop (2009). *http://hadoop. apache. org*

An, L., Linderman, M., Qi, J., Shortridge, A., & Liu, J. (2005). Exploring complexity in a human-environment system: An agent-based spatial model for multidisciplinary and multiscale integration. *Annals of the Association of American Geographers, 95*(1), 54-79. doi:10.1111/j.1467-8306.2005.00450.x

An, L. (2012). Modeling human decisions in coupled human and natural systems: Review of agent-based models. *Ecological Modelling, 229*, 25-36. doi:10.1016/j.ecolmodel.2011.07.010

Amdahl, G. (1967). Validity of the Single Processor Approach to Achieving Large-Scale Computing Capabilities. AFIPS Conference Proceedings (30), 483–485.

Barrow, C. J. (1998). River basin development planning and management: A critical review. *World Development, 26*(1), 171-186. doi:10.1016/S0305-750X(97)10017-1

Becu, N., Perez, P., Walker, A., Barreteau, O., & Le Page, C. (2003). Agent based simulation of a small catchment water management in northern thailand description of the CATCHSCAPE model RID F-1968-2010 RID E-6856-2010. *Ecological Modelling, 170*(2-3), 319-331. doi:10.1016/S0304-3800(03)00236-9

Bennett, D. A., & Tang, W. (2006). Modelling adaptive, spatially aware, and mobile agents: Elk migration in yellowstone. *International Journal of Geographical Information Science,20*(9), 1039-1066. doi:10.1080/13658810600830806

Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences of the United States of America, 99,* 7280-7287. doi:10.1073/pnas.082080899

Borthakur, D. (2007). Hadoop distributed file system. *Apache Software Foundation*.

Curbera, F., Duftler, M., Khalaf, R., Nagy, W., Mukhi, N., & Weerawarana, S. (2002). Unraveling the web services web: an introduction to SOAP, WSDL, and UDDI. *IEEE Internet Computing 6* (2), 86-93.

Castronova, A. M., Goodall, J. L., & Elag, M. M. (2013). Models as web services using the open geospatial consortium (OGC) web processing service (WPS) standard.*Environmental Modelling & Software, 41*, 72-83. doi:10.1016/j.envsoft.2012.11.010

Da-Jun, T., Tang, F., Lee, T., Sarda, D., Krishnan, A., & Goryachev, A. (2005). Parallel computing platform for the agent-based modeling of multicellular biological systems. *Parallel and distributed computing: Applications and technologies* (pp. 5-8) Springer.

Dean, J., & Ghemawat, S. (2008). Mapreduce: Simplified data processing on large clusters. *Communications of the ACM, 51*(1), 107-113. doi:10.1145/1327452.1327492

Deissenberg, C., van der Hoog, S., & Dawid, H. (2008). EURACE: A massively parallel agent-based model of the european economy. *Applied Mathematics and Computation, 204*(2), 541-552. doi:10.1016/j.amc.2008.05.116

Epstein, J.M. (1999). Agent-Based Computational Models and Generative Social Science. *Complexity,* 4, 41–60.

Feuillette, S., Bousquet, F., & Le Goulven, P. (2003). SINUSE: A multi-agent model to negotiate water demand management on a free access water table. *Environmental Modelling & Software, 18*(5), 413-427. doi:10.1016/S1364-8152(03)00006-9

Gilbert, G.N., & Troitzsch, K. G. (1999). Simulation for the Social Scientist, Milton Keynes, Open University Press, UK.

Goetz, B., Peierls, T., Bloch, J., Bowbeer, J., Holmes, D., & Lea, D. (2006). Java Concurrency in Practice, Addison-Wesley.

Gong, Z., Tang, W., Bennett, D. A., & Thill, J. (2013). Parallel agent-based simulation of individual-level spatial interactions within a multicore computing environment. *International*

Journal of Geographical Information Science, 27(6), 1152-1170. doi:10.1080/13658816.2012.741240

Goodall, J. L., Robinson, B. F., & Castronova, A. M. (2011). Modeling water resource systems using a service-oriented computing paradigm. *Environmental Modelling & Software,26*(5), 573-582. doi:10.1016/j.envsoft.2010.11.013

Granell, C., Diaz, L., & Gould, M. (2010). Service-oriented applications for environmental models: Reusable geospatial services. *Environmental Modelling & Software, 25*(2), 182-198. doi:10.1016/j.envsoft.2009.08.005

Haefner, J. W. (1992). In Deangelis D. G.,LJ (Ed.), *Parallel computers and individual-based models - an overview*

Hadoop Map/Reduce tutorial (2013). https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html

Hargreaves, G. H., & Samani, Z. A. (1985). Reference crop evapotranspiration from ambient air temperature. *American Society of Agricultural Engineers, 1*(2), 96-99

Hipel, K. W., Jamshidi, M. M., Tien, J. M., & White,Chelsea C.,,III. (2007). The future of systems, man, and cybernetics: Application domains and research methods. *IEEE Transactions on Systems Man and Cybernetics Part C-Applications and Reviews, 37*(5), 726-743. doi:10.1109/TSMCC.2007.900671

Huhns, M., & Singh, M. (2005). Service-oriented computing: key concepts and principles. *IEEE Internet Computing 9* (1), 75-81.

Hunt, R. J., Luchette, J., Schreuder, W. A., Rumbaugh, J. O., Doherty, J., Tonkin, M. J., & Rumbaugh, D. B. (2010). Using a cloud to replenish parched groundwater modeling efforts. *Ground water, 48*(3), 360-365.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Keutzer, K., & Mattson, T. (2009, June). Our Pattern Language (OPL): A design pattern language for engineering (parallel) software. *In ParaPLoP Workshop on Parallel Programming Patterns*.

Lansing, J. S. (1999). Anti-Chaos, Common Property and the Emergence of Cooperation, in Dynamics in Human and Primate Societies: Agent-Based Modelling of Social and Spatial Processes, ed. Timothy Kohler and George Gumerman, Santa Fe Institute and Oxford University Press: 207-224.

Lee, P. M. (2004). *Bayesian statistics: An introduction*. Arnold Publishing, 2004. Third edition.

Liu, J., Dietz, T., Carpenter, S. R., Alberti, M., Folke, C., Moran, E., . . . Taylor, W. W. (2007). Complexity of coupled human and natural systems. *Science, 317*(5844), 1513-1516. doi:10.1126/science.1144004

Liu, Y., Sun, A. Y., Nelson, K., & Hipke, W. E. (2013). Cloud computing for integrated stochastic groundwater uncertainty analysis. *International Journal of Digital Earth, 6*(4), 313-337

Massingill, B. L., Mattson, T. G., & Sanders, B. A. (2005). Reengineering for Parallelism: An Entry Point into PLPP (Pattern Language for Parallel Programming) for Legacy Applications. *Proceedings of the Twelfth Pattern Languages of Programs Workshop (PLoP 2005)*.

Mattson, T. G., Sanders B. A., & Massingill, B. L. (2004). A Pattern Language for Parallel Programming. Addison Wesley Software Patterns Series.

North M. J., & Macal, C.M. (2007). Managing Business Complexity: Discovering Strategic Solutions with Agent-Based Modeling and Simulation. Oxford University Press

McKusick, V. (2003). Final report for the special master with certificate of adoption of rrca groundwater model. *State of Kansas v. State of Nebraska and State of Colorado, in the Supreme Court of the United States, 3605*.

Monticino, M., Acevedo, M., Callicott, B., Cogdill, T., & Lindquist, C. (2007). Coupled human and natural systems: A multi-agent-based approach. *Environmental Modelling & Software, 22*(5), 656-663. doi:10.1016/j.envsoft.2005.12.017

Muller, P.A. (1997). Modélisation objet avec UML, Eyrolles Paris.

Mulvey, J. M., Vanderbei, R. J., & Zenios, S. A. (1995). Robust optimization of large-scale systems. *Operations Research, 43*(2), 264-281. doi:10.1287/opre.43.2.264

Murphy, K. P. (2007). Conjugate Bayesian analysis of the Gaussian distribution. *Technical report*, University of British Columbia.

Nielsen, M. (2009). Write your first MapReduce program in 20 minutes. Michael's main blog: http://michaelnielsen.org/blog/write-your-first-mapreduce-program-in-20-minutes/

Pahl-Wostl, C. (2005). Information, public empowerment, and the management of urban watersheds. *Environmental Modelling & Software, 20*(4), 457-467. doi:10.1016/j.envsoft.2004.02.005

Palazzo, A. M. (2009). Farm-Level Impacts of Alternative Spatial Water Management Policies for the Protection of Instream Flows. *Master Thesis*, University of Illinois at Urbana-Champaign.

Papajorgji, P., Beck, H. W., & Braga, J. L. (2004). An architecture for developing service-oriented and component-based environmental models. *Ecological Modelling, 179*(1), 61-76. doi:10.1016/j.ecolmodel.2004.05.013

Parker, J., & Epstein, J. M. (2011). A distributed platform for global-scale agent-based models of disease transmission. *ACM Transactions on Modeling and Computer Simulation (TOMACS), 22*(1), 2.

Parker, D. C., Manson, S. M., Janssen, M. A., Hoffmann, M. J., & Deadman, P. (2003). Multi-agent systems for the simulation of land-use and land-cover change: A review. *Annals of the Association of American Geographers, 93*(2), 314-337. doi:10.1111/1467-8306.9302004

Poo, D., Kiong, D., & Ashok, S. (2007). Java Database Connectivity. *Object-Oriented Programming and Java*, 297-314.

Reeves, H. W., & Zellner, M. L. (2010). Linking MODFLOW with an agent-based land-use model to support decision making. *Ground Water, 48*(5), 649-660. doi:10.1111/j.1745-6584.2010.00677.x

Reynolds, R. G., Kobti, Z., Kohler, T. A., & Yap, L. Y. L. (2005). Unraveling ancient mysteries: Reimagining the past using evolutionary computation in a complex gaming environment. *IEEE Transactions on Evolutionary Computation, 9*(6), 707-720. doi:10.1109/TECV.2005.856206

Robinson, D. T., & Brown, D. G. (2009). Evaluating the effects of land-use development policies on ex-urban forest cover: An integrated agent-based GIS approach. *International Journal of Geographical Information Science, 23*(9), 1211-1232. doi:10.1080/13658810802344101

Snir, M., Otto, S., Huss-Lederman, S., Walker, D., and Dongarra, J. (1998). MPI-The Complete Reference: Vol. 1. *The MPI Core (2nd ed.)*. The MIT Press

Tang, W., & Bennett, D. A. (2012). Reprint of: Parallel agent-based modeling of spatial opinion diffusion accelerated using graphics processing units. *Ecological Modelling, 229*, 108-118.

Tang, W., Wang, S., Bennett, D. A., & Liu, Y. (2011). Agent-based modeling within a cyberinfrastructure environment: a service-oriented computing approach. *International Journal of Geographical Information Science*, *25*(9), 1323-1346.

Tze Ling Ng, Eheart, J. W., Cai, X., & Braden, J. B. (2011). An agent-based model of farmer decision-making and water quality impacts at the watershed scale under markets for carbon allowances and a second-generation biofuel crop. *Water Resources Research, 47,* W09519. doi:10.1029/2011WR010399

USDA (1967). Irrigation water requirements. *Tech. Release No. 21*. United States Dept. of Agr., Soil Conservation Service, Washington, D.C

Wächter, A., & Biegler, L. (2006). On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming, 106*(1), 25-57. doi:10.1007/s10107-004-0559-y

Zhang, K., Kimball, J. S., Mu, Q., Jones, L. A., Goetz, S. J., & Running, S. W. (2009). Satellite based analysis of northern ET trends and associated changes in the regional water balance from 1983 to 2005. *Journal of Hydrology, 379*(1), 92-110.

Zhang, K., Kimball, J. S., Nemani, R. R., & Running, S. W. (2010). A continuous satellite-derived global record of land surface evapotranspiration from 1983 to 2006.*Water Resources Research, 46*(9)

3 CHAPTER III

Global sensitivity analysis for the coupled agent-based model and groundwater model using Hadoop

3.1 Introduction

Society faces a number of water challenges due to human decisions, inducing water scarcity, waste, pollution and unsustainable management. Incorporating human factors into water resources systems analysis provides a framework to address sustainability issues in water resources planning and management (Molle, 2009). This framework will allow us to understand the interactions between human and environmental systems while tackling the challenging water issues arising from human activities. The coupling component models (CCMs) approach (Kelly et al., 2013) is applied to designing a socio-hydrological model by coupling a multi-agent system (MAS) behavioral model with a physically-based groundwater model- the Republican River Compact Administration (RRCA) model. The MAS model describes farmers' behaviors in terms of groundwater pumping decisions; the RRCA model simulates the water level in an aquifer and water exchange between the aquifer and streams (RRCA, 2003; Mulligan et al., 2014). Thus, the coupled socio-hydrological model enables us to have a quantitative understanding of the impacts of the farmers' pumping behaviors on the water table and the baseflow requirement for ecosystems, and vice versa (Hu et al., 2015).

However, in behavioral modeling and simulation (Van Hemel et al., 2008), virtually all simulations include some random elements in both their initial conditions and their mechanisms for change (Axelrod, 1997). Human behavioral parameters as inputs to these models are highly uncertain and variable. The input data are either not directly available or may only be indirectly inferred from related information, and their uncertainty and error must be considered (Liebl, 1995; Bier, 2011). In addition, as in many cases in the social sciences, the cause and effect relations in the systems of interest are not yet well understood. A single run of the simulation model with one given set of parameters will not be able to represent the real underlying uncertainties, and the result can be misleading (Axelrod, 1997). By changing the input systematically, sensitivity analysis can help explore the relationships and mechanisms that are not yet well understood, reveal the possible variations in the results, and highlight the most important processes especially in social sciences

(Chattoe, et al., 2000). For example, Happe (2005) presents the application of sensitivity analysis with an agent-based model named AgriPolis in order to identify the factors which most affect average economic land rent per hectare.

With simple simulation models, sensitivity analysis is often conducted by varying one parameter at a time while keeping the remaining parameters constant. For complex simulation models, such as coupled socio-hydrological models, this 'One-At-a-Time' (OAT) approach ignores the possible interactions between input parameters, and therefore is not able to capture their impacts on the model outputs as the result of their interactions (Kleijnen et al., 2003; Happe 2005). Derivative-based local sensitivity analysis is often limited to a few simulation models with analytical solutions of the derivatives with respect to the parameters of interest (Huard and Mailhot, 2006). It becomes unwarranted when the inputs are uncertain, and less informative, in particular when we want to explore the rest of the space of the input factors, other than the base points where the derivatives are calculated (Saltelli et al., 2008). Meanwhile, we are aware that the quantities of interest for social scientists usually lie in the effects of input variables on the spatio-temporal evolution of output variables. However, those sensitivity analysis methods lack of the ability to analyze the spatial and temporal effects of human behavioral uncertainty, which are of great importance to understanding the interactions between human and hydrological systems.

Compared to local sensitivity analysis, Wainwright et al., (2014) claim that global sensitivity analysis (GSA) can provide robust sensitivity measures in the presence of nonlinearity and interactions among the parameters. However, GSA can be computationally expensive using Monte Carlo methods, since it usually requires a large number of model evaluations (Sobol, 1993; Saltelli, 2002; Saltelli et al., 2008). Thus, in practice, conducting the Monte Carlo-based GSA with complex models can be infeasible. In this chapter, we propose a methodological framework for GSA applied to large scale socio-hydrological models by combining a Hadoop-based cloud computing approach for model evaluation, with a Polynomial Chaos Expansion (PCE) based variance decomposition approach for estimation of the sensitivity indices. We will demonstrate how these techniques make GSA computationally tractable for complex socio-hydrological models. The rest of the chapter is organized as follows. We start with a brief introduction of our coupled socio-hydrological models. Then, we address two major challenges arising from GSA with large-scale socio-hydrological models: 1) the computational cost associated with running the computationally intensive coupled socio-hydrological models to generate sufficient model outputs for GSA; 2) sensitivity analysis methods that can effectively utilize a large amount of multidimensional data as the result of Monte Carlo runs and capture the spatial and temporal variations of input variables on model outputs. Finally, we will conclude with the advantages of the proposed computational framework for large-scale socio-hydrological models.

3.2 Background

3.2.1 Case Study Site

The Republican River originates in the high plains of northeastern Colorado, western Kansas and southern Nebraska. The basin covers approximately 25,018 square miles (~16 million acres) of the three states, and is encompassed by the underlying High Plains aquifer as shown by Figure 3-1. Due to the intensive agriculture development in the Republican River Basin since the 1970s, there has been a significant increase of groundwater use for irrigation. Water conflicts and lawsuits arise from the sharing of the groundwater resources among the three states in the Republican River Basin: Colorado, Kansas and Nebraska. As part of the US Supreme Court settlement, a comprehensive groundwater model, the Republican River Compact Administration (RRCA) groundwater model which uses MODFLOW-2000 with additional modules, was developed through the collaboration of the three affected states, the U.S. Geological Survey, and the U.S. Bureau of Reclamation (McKusick, 2003). Using the principle of water balance, the RRCA model, which allows for spatial variability in hydraulic conductivity (K), evapotranspiration (ET), recharge, etc. is used to represent groundwater flow in the Republican River Basin and determine the time, location and amount of stream depletions as the result of well pumping (RRCA, 2003; Mulligan et al., 2014).

3.2.2 Coupled MAS Model and RRCA Model

The multi-agent system is characterized as a collection of autonomous decision-making and interactive entities, namely agents. These agents are autonomous, interdependent and adaptive, and they follow a base-level set of behavioral rules. Those rules can be altered by other high-level sets of rules for agents to learn and adapt to the environment (North and Macal, 2007). The design of the MAS model follows a bottom-up approach to assist in the spatial-temporal exchange of information. In this study, we developed a multi-agent system (MAS) model to describe farmers' decision-making processes on groundwater pumping for irrigation in this region by taking various environmental and socioeconomic factors into account, and coupled it with the RRCA model to simulate the interactions between farmers' pumping behaviors and the groundwater system as shown by Figure 3-2.

For individual components of the coupled models, they often work on different space and time scales and necessary disaggregation and aggregation procedures are required to couple them together (Kelly et al., 2013). For example, in these coupled socio-hydrological models, each agent is defined as a county within the High Plains aquifer as shown by Figure 3-1 and characterized by the five behavioral parameters (κ_{pr}, ν_{pr}, κ_{prep}, ν_{prep} and λ) in Table 3-1 (i.e., 46 agents and 230 parameters in total). For parameters κ_{pr}, ν_{pr}, κ_{prep} and ν_{prep}, the larger the parameter values, the more confidence the agents have on the prior knowledge of the mean and variance of the crop prices and precipitation. For parameter λ, the larger the value, the more cautious the agents are to take the risk in pursuit of higher crop profit return. Given the behavioral parameters, each agent makes annual predictions of the future crop prices and precipitation via Bayesian learning (See Appendix B.1). The estimated crop prices and precipitation are then fed into the stochastic utility maximization model which mimics agent's decisions on the choices of crop types, the corresponding planted irrigated and rainfed crop areas and the annual groundwater usage (Hu et al., 2015). The annual groundwater withdrawal is then converted to the monthly pumping rate for the wells (shown as red dots in Figure 3-1) to drive the RRCA model. The outputs of the RRCA model are used as the feedback to the MAS model for the next year. One of the key feedbacks is the water table, which is used to evaluate the impacts of agents' pumping decisions on groundwater. The water table is converted to the depth to groundwater and then translated into the crop revenue function in terms of energy cost, which affects agents' decisions on pumping in the following year (Hu et al., 2015).

Given the complexity of the underlying models and their links, it is very challenging to fully understand the true uncertainty in the coupled models (Kelly et al., 2013). In this study, we assume that the parameters in the RRCA model are constant and the only uncertainty is in the MAS model as the result of the variations of the behavioral parameters. We recognize that there can be uncertainty in the hydrogeological parameters in the RRCA model, but we leave that for future investigations. Some test results have shown the impacts of the behavioral parameters on the selection of crops, irrigation area, crop profits and groundwater usage (Hu et al., 2015). Through sensitivity analysis, we want to identify which parameter(s) have the most significant impacts on the groundwater table. As a result, it can help us gain some insights into the causes and effects between farmers' pumping behaviors and groundwater decline.

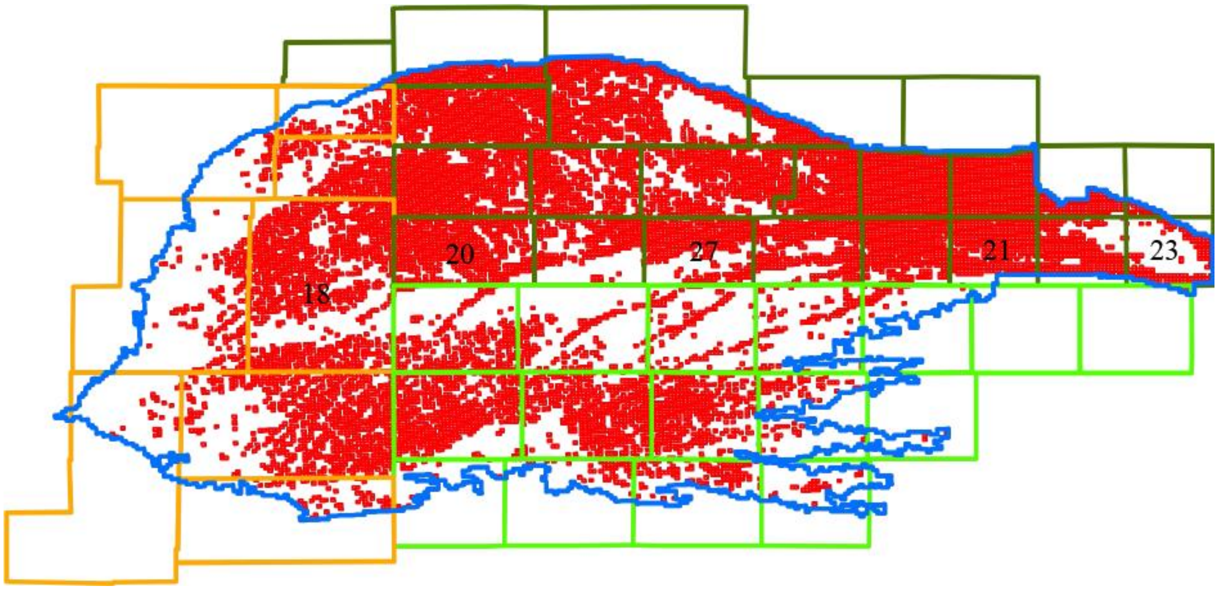


Figure 3-1 The aerial view of the pumping wells (red dots) and High Plains aquifer (blue line) in MODFLOW-2000 and the overlapping counties (blocks) of different states (orange: Colorado; light green: Kansas; spruce green: Nebraska; each county is treated as an agent and the numbers are selected agent IDs).

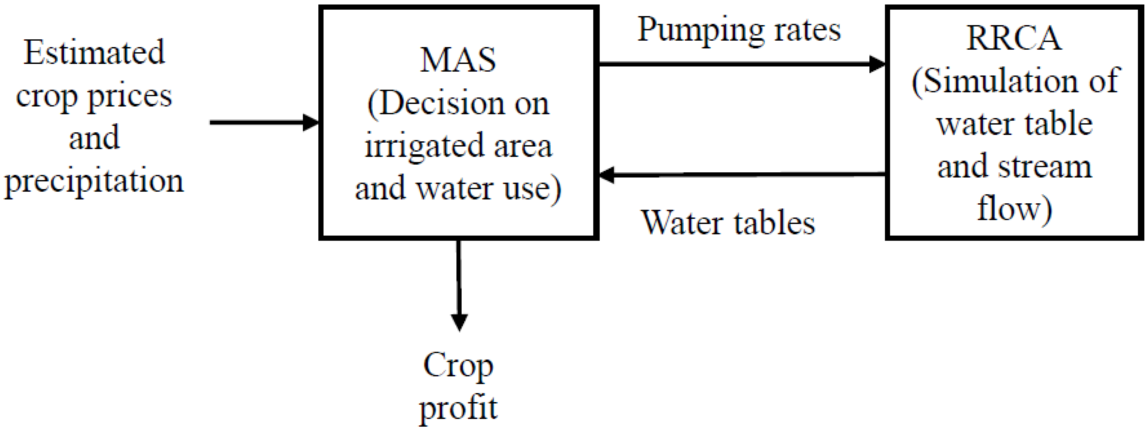

Estimated crop prices and precipitation
MAS (Decision on irrigated area and water use)
Pumping rates
Water tables
RRCA (Simulation of water table and stream flow)
Crop profit

Figure 3-2 Coupling of the MAS model (with an annual time interval) and RRCA model (with monthly stress period). The simulation period is from year 1993 to 2006.

Table 3-1 Five behavioral parameters (κ_{pr} , ν_{pr} , κ_{prep} , ν_{prep} and λ). It is assumed that all parameters are independent and follow uniform distribution with different ranges as shown in the table.

Parameter	Range	Definition
κ_{pr}	[0.5,5]	agents' beliefs in their prior knowledge of the mean of crop prices.
ν_{pr}	[5,50]	agents' beliefs in their prior knowledge of the variance of crop prices.
κ_{prep}	[0.5,5]	agents' beliefs in their prior knowledge of the mean of precipitation.
ν_{prep}	[5,50]	agents' beliefs in their prior knowledge of the variance of precipitation.
λ	[0,20]	agents' attitudes towards the fluctuations of crop profits.

3.3 Methodology

In the following sections, we first discuss the sampling approach that can generate sparse but well-represented samples from the parameter domain. Then, we delve into the computational issues which result from running a large number of simulations of the coupled socio-hydrological model with the samples from the previous step. Once we obtain the large amount of model outputs, we introduce an efficient approach to estimate sensitivity indices.

3.3.1 Sampling Generation

The five behavioral parameters defined for each agent describe the agent's preference between the prior knowledge and the historical experience of crop prices and precipitation when predicting their values during the crop planting season (Hu et al., 2015). However, no data is available to measure the correlation between agents' preferences across county lines; in the current study, we assume all the parameters are independent and follow a uniform distribution with different ranges shown in Table 3-1 and leave the case of the correlation between agents' preferences for future investigation. The coupled Latin Hypercube Sampling (LHS) with a Genetic Algorithm (Stocki, 2005), called *geneticLHS* in R (R Package 'lhs', Carnell and Carnell, 2012) is applied for sampling the parameter sets that are expected to well represent the entire parameter domain through maximizing the mean distance from each design point to all the other points in the domain. Thus, the designed points are spread out as much as possible. The sampling method works as follows. First, we apply the *geneticLHS* to generate a large number of sample sets as the candidate input sets for the MAS model, and each sample set contains the values of the five behavioral parameters. For each run of the coupled MAS model and RRCA model (i.e. running the coupled models over 14 years from 1993 to 2006), we randomly choose one input set from the candidate input sets, assign it to one individual agent and repeat the procedure for all 46 agents. Then, we iterate the run over $\boldsymbol{N}$ times and thus generate a data set, including $\boldsymbol{N}$ independent and identically distributed samples for all 230 parameters in the coupled models denoted by $\boldsymbol{S} = \{\boldsymbol{X}^{(1)},...,\boldsymbol{X}^{(\mathrm{N})}\}$, where $\boldsymbol{X}^{(i)} = (\boldsymbol{X}_1^{(i)},...,\boldsymbol{X}_{230}^{(i)})$ is treated as one model input, and hence one scenario for sensitivity analysis is thus defined as a single execution of the coupled models with $\boldsymbol{X}^{(i)}$.

3.3.2 Cloud Computing with Hadoop

As mentioned above, each agent is characterized by five human behavioral parameters (i.e. 230 independent parameters in total for all 46 agents in the coupled models). It is expected to run the coupled models at least twice as many times as the total number of parameters (i.e. 460 times) in order to estimate the effect of changing each parameter when using the OAT approach (Saltelli, et al., 2008). In the case of GSA with a variance decomposition approach accounting for the impact on model outputs incurred by parameter interactions, even more model evaluations are desired. However, one single sequential execution of the coupled models over 14 years takes one hour on a desktop machine (2.4GHz Dual Core; Hu et al., 2015). For example, if we want to run the coupled models 1,000 times for sensitivity analysis (i.e., 1,000 scenarios), the total computation time is approximately 42 days. Here, we assumed that the interactions between agents are not implemented explicitly, but implicitly through the RRCA model and much of their computation work can thus be executed independently without the need of message passing among agents (Hu et al., 2015), and each scenario for sensitivity analysis is independent. These features make sensitivity analysis with the coupled models well suited for parallel computing. Cloud computing, as one kind of parallel computing, is considered as a cost-saving means to bring this unprecedented computing power, and also has a great advantage of on-demand access in comparison with conventional supercomputers. Hunt et al. (2010) point out that sensitivity analysis or auto-calibration with the complex models can benefit from cloud computing techniques, among which Apache Hadoop is a commonly used open-source software framework for a large amount of data storage and processing on clusters of commodity hardware. This framework consists of two main components: Hadoop MapReduce framework (Dean and Ghemawat, 2008) and Hadoop

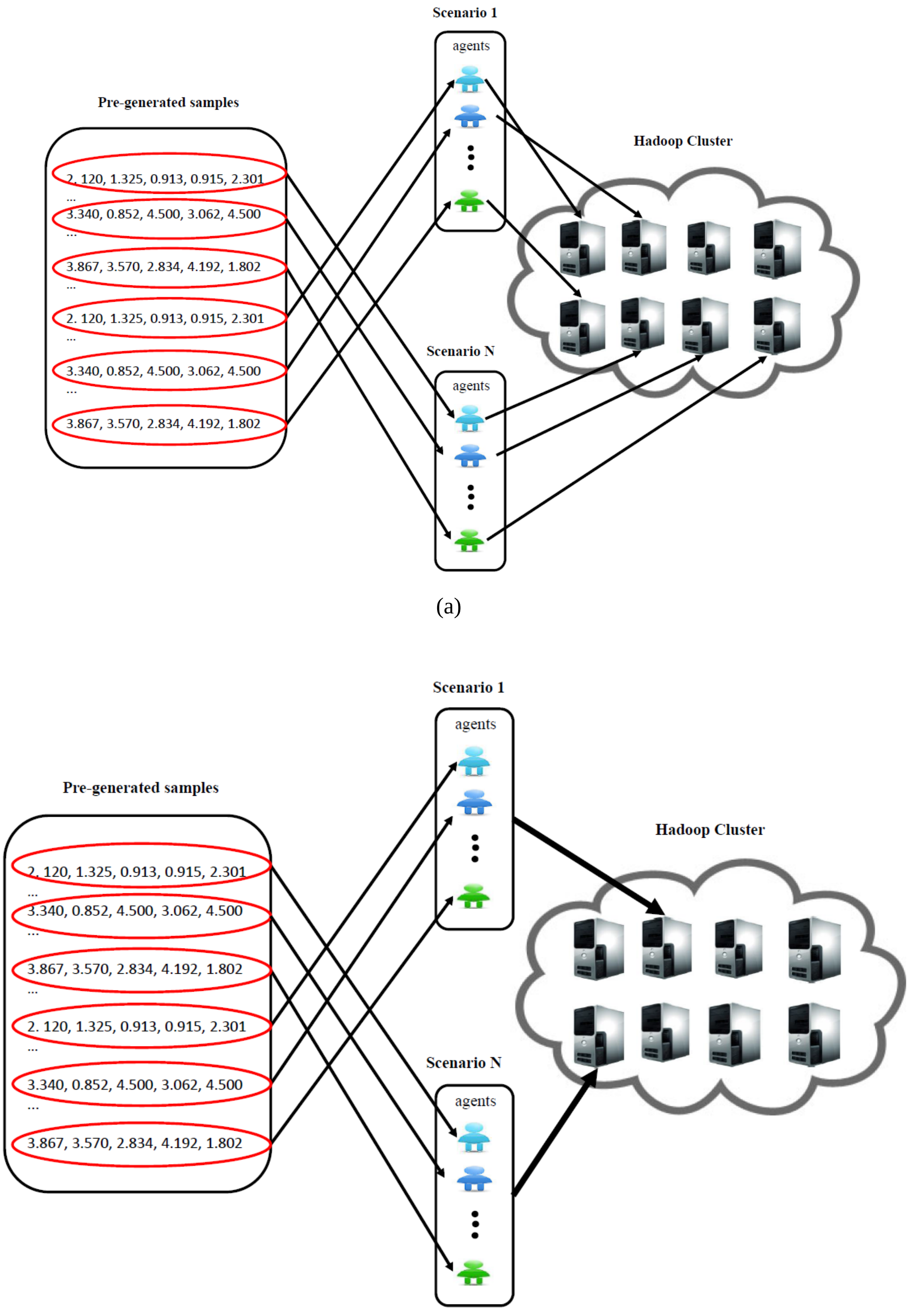
Distributed File Systems (HDFS) (Borthakur, 2007). MapReduce framework includes two phases, map phase and reduce phase. A MapReduce job usually splits the input dataset into independent sub-datasets on different nodes, processes them in the map phase and outputs the results in terms of key-value pairs in a completely parallel manner (Hadoop Map/Reduce tutorial, 2013). The framework then shuffles and sorts the key-value pairs according to the keys and uses them as the inputs to the reduce phase. In the reduce phase, the key-value pairs will be combined together to form a smaller set of values given the same keys. From a programmer's point of view, MapReduce, similar to other Java libraries, is imported at the beginning of a program. One only needs to implement the map and reduce function defined in the map and reduce phase, and pass the input data into these functions (Nielsen, 2009). HDFS can be used to store both the input and the output files of the job. We develop two approaches using the Apache Hadoop framework to address the computational issues arising from the sensitivity analysis with our socio-hydrological model.

3.3.2.1 Approach I: Running different agents with different machine nodes

Given the assumption that agents only have implicit interactions through the RRCA model, the first approach is developed to execute the tasks associated with the individual agent in parallel during the map phase. For each scenario, each of the 46 agents randomly chooses values for the five target behavioral parameters from the pre-generated samples using the *geneticLHS* as shown by Figure 3-3 (a), and executes their tasks. The output in the map phase is the key-value pair associated with the individual agent, where the scenario ID is the key and the pumping rates are the value. We repeat the process over ***N*** times in the map phase. Notice that available machine nodes are randomly allocated for different agents to execute their tasks. In the reduce phase, all the key-value pairs with the same scenario IDs are grouped together and the values are used as the input for the RRCA model. After the execution of the RRCA model in the reduce phase completes, the output, such as the water table is saved to the HDFS and used as the input for the MAS model in the consecutive year. The coupled models require a number of read-only files which are stored in the distributed cache. Figure 3-4 (a) shows the integration of the Apache Hadoop framework into the coupled models. Each map/reduce loop represents one-year execution of the coupled models. The total simulation period of the coupled models is 14 years for the case study as described by the pseudo-code of Algorithm I. The input sets of the behavior parameters for a specific agent and the corresponding outputs of the groundwater table under ***N*** scenarios are then used for sensitivity analysis. It is noted that this approach requires significant overhead associated with starting a MapReduce job, and agents also require a very large number of input files that must be distributed locally to each task in the map phase (map task). In order to amortize the computational cost of copying the input files to the map tasks, we need to estimate the amount of time that it takes to execute the tasks for a single agent, and the amount of time to copy the input files to the map task and start the map task. To find the optimal number of tasks, the total computation time to execute the ***N*** simulations in the map phase is minimized as shown in the following optimization model:

$$\begin{aligned}
&\mathit{Minimize}\,(c+O+t\cdot n)\cdot m\\
&\mathit{Subject\ to}\ t\cdot n\geq c+O;\\
&\qquad\qquad n\cdot m=T;\\
&\qquad\qquad n\geq 1
\end{aligned}\tag{1}$$

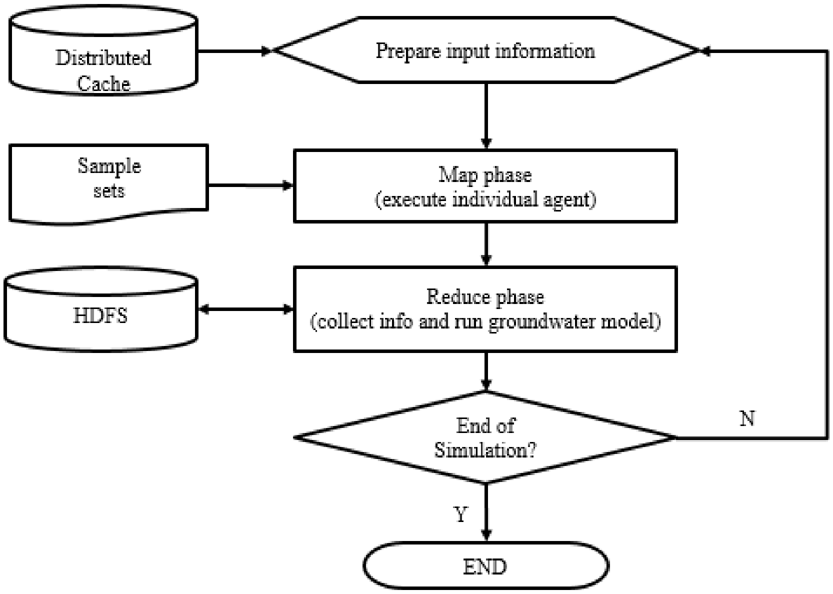
where *c* is the time to copy the input files for the new map task; *O* is the overhead associated with initializing a new MapReduce job; *t* is the amount of time it takes to execute the tasks for a single agent; *n* is the number of agents that run in a single map task and *m* is the number of map tasks spawned; *T* is the total number of agents to be executed for sensitivity analysis, that is, ***46·N***. As shown by the first constraint, it is worthwhile to initiate a new MapReduce job only when the total amount of time to execute the task for *n* agents is larger than the time to prepare the input files for the new MapReduce job and the associated overhead to initialize the MapReduce job. Otherwise, we can distribute the *n* agents to other existing MapReduce jobs.



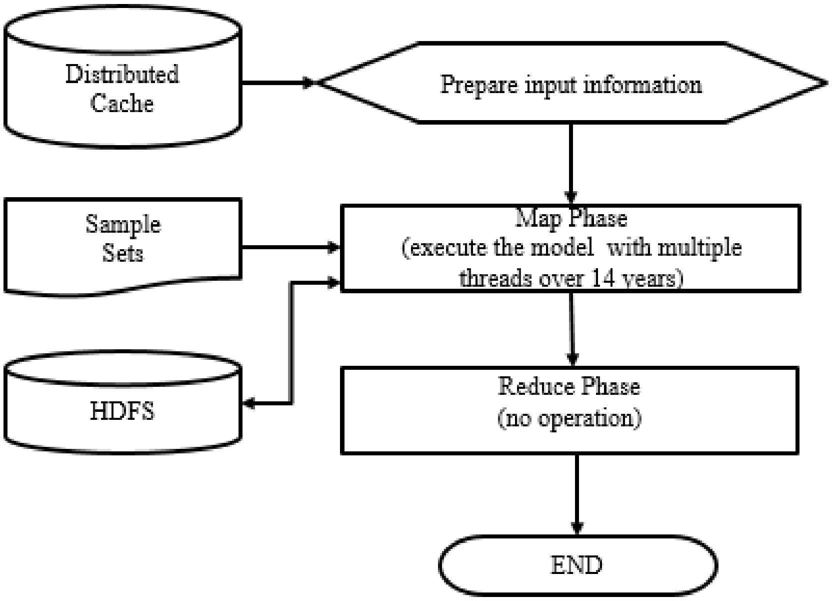
(a)

(b)

Figure 3-3 (a) schematic diagram of Approach I: agents are randomly chosen by different map tasks initiated at different machine nodes and (b) schematic diagram of Approach II: agents are lumped together and different scenarios are distributed by the MapReduce framework and executed at different machine nodes.

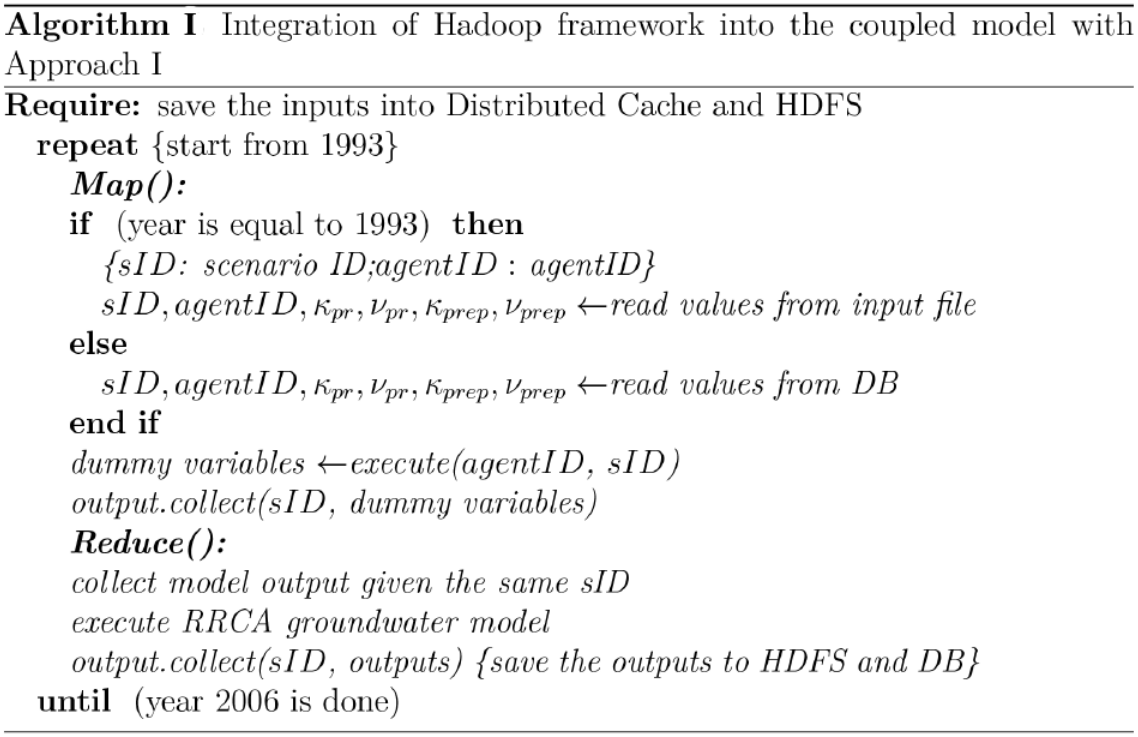

Distributed Cache
Prepare input information
Sample sets
Map phase (execute individual agent)
HDFS
Reduce phase (collect info and run groundwater model)
End of Simulation?
N
Y
END

(a)


Distributed Cache
Prepare input information
Sample Sets
Map Phase (execute the model with multiple threads over 14 years)
HDFS
Reduce Phase (no operation)
END

(b)

Figure 3-4 (a) Approach I: Integration of the Apache Hadoop framework into the coupled MAS and RRCA model. A number of read-only files for the coupled models are stored in the distributed cache. HDFS is used to save the model outputs. Agents are spread over and executed in different map tasks initiated at different nodes, and all the key-value pairs with the same scenario IDs are grouped together, where the values are used as the input for the RRCA model. This loop continues over the entire simulation period. (b) Approach II: Integration of the Apache Hadoop framework into the coupled MAS and RRCA models. The single simulation of the coupled models over the entire simulation period is executed only in the map phase.



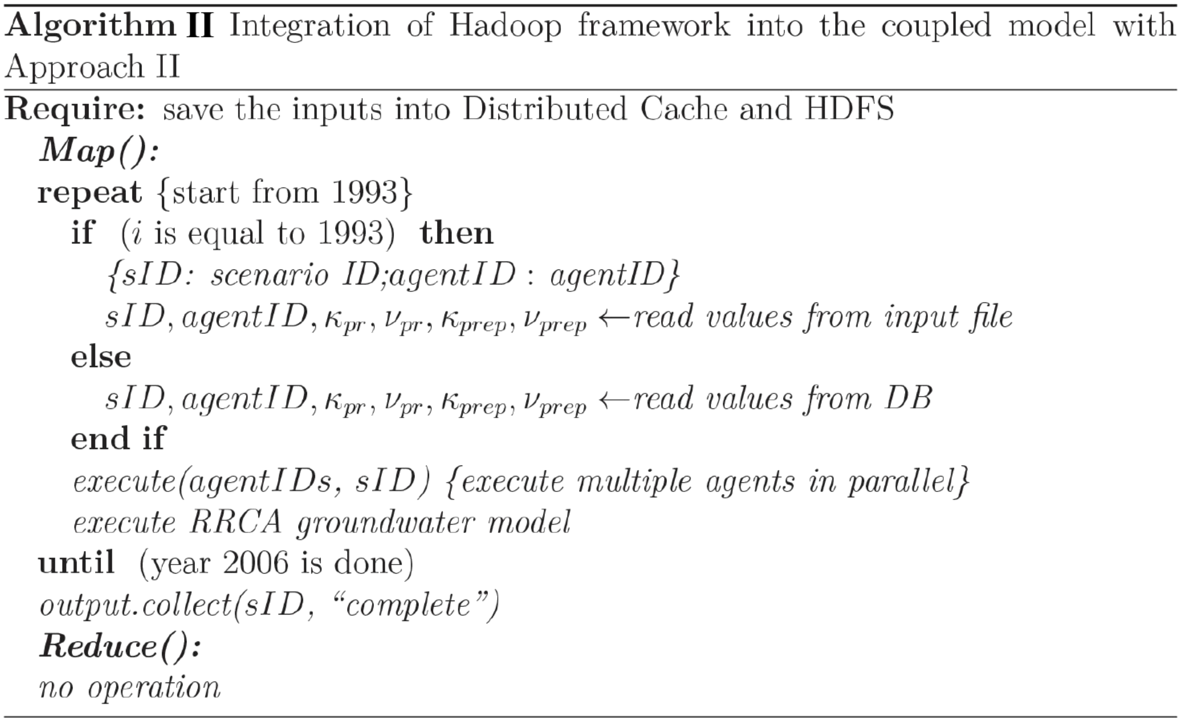
Algorithm I Integration of Hadoop framework into the coupled model with Approach I

```
Require: save the inputs into Distributed Cache and HDFS
  repeat {start from 1993}
    Map():
    if (year is equal to 1993) then
      {sID: scenario ID;agentID : agentID}
      $sID, agentID, \kappa_{pr}, \nu_{pr}, \kappa_{prep}, \nu_{prep} \leftarrow$ read values from input file
    else
      $sID, agentID, \kappa_{pr}, \nu_{pr}, \kappa_{prep}, \nu_{prep} \leftarrow$ read values from DB
    end if
    dummy variables ←execute(agentID, sID)
    output.collect(sID, dummy variables)
    Reduce():
    collect model output given the same sID
    execute RRCA groundwater model
    output.collect(sID, outputs) {save the outputs to HDFS and DB}
  until (year 2006 is done)
```

Algorithm I: Implementation of Approach I: integration of the Apache Hadoop framework into the coupled models (DB: database).

3.3.2.2 Approach II: Running different scenarios with different machine nodes

Different from the first approach where agents are randomly chosen by different map tasks initiated at different machine nodes and the key-value pairs of results are collected in the reduce phase, the second approach executes the single simulation of the coupled models over the entire simulation period within the map phase. In this sense, all 46 agents are lumped together and execute their tasks in the same map task. It works as follows: similar to Approach I, each of the 46 agents randomly chooses values for the five behavioral parameters from the pre-generated samples as the inputs for the MAS model as shown by Figure 3-3 (b). Since there exists no explicit communication between agents for the MAS model, the MAS model can then run in parallel with multithreading techniques in the map phase (Hu et al., 2015). The outputs from the MAS model are then used as the inputs for the RRCA model, and it iterates over the next year until the end of the entire simulation period as shown by Figure 3-4 (b) (See Algorithm II regarding the implementation of this approach). In addition, note that in Approach II different scenarios are distributed by the MapReduce framework and executed over the available nodes as shown by Figure 3-3 (b), rather than all agents from all ***N*** scenarios for every year spreading over different nodes in the first approach as shown by Figure 3-3 (a).



Algorithm II Integration of Hadoop framework into the coupled model with Approach II

```
Require: save the inputs into Distributed Cache and HDFS
  Map():
  repeat {start from 1993}
    if ($i$ is equal to 1993) then
      {sID: scenario ID;agentID : agentID}
      $sID, agentID, \kappa_{pr}, \nu_{pr}, \kappa_{prep}, \nu_{prep} \leftarrow$ read values from input file
    else
      $sID, agentID, \kappa_{pr}, \nu_{pr}, \kappa_{prep}, \nu_{prep} \leftarrow$ read values from DB
    end if
    execute(agentIDs, sID) {execute multiple agents in parallel}
    execute RRCA groundwater model
  until (year 2006 is done)
  output.collect(sID, "complete")
  Reduce():
  no operation
```

Algorithm II: Implementation of Approach II: Integration of the Apache Hadoop framework into the coupled models.

3.3.3 Global Sensitivity Analysis

In the first part of the chapter we develop two approaches based on the Apache Hadoop framework to address the computational issues arising from the model evaluations with the coupled socio-hydrological models. As mentioned above, these model inputs and outputs from model evaluations are going to be used for GSA. In the following sections, we will discuss a promising GSA approach that has the potential to handle the large amount of multidimensional data that results from the large number of model evaluations, and explore the spatio-temporal variations of the input behavioral parameters on the model outputs.

GSA has been widely used in hydrology and environmental fields (Francos et al., 2003; Pappenberger et al., 2008; Moreau et al., 2013; Zhang et al., 2013; Garcia-Cabrejo and Valocchi, 2014; Sweetapple et al., 2014, 2013). The quantitative approach to GSA that has been used extensively in recent years is called variance or Sobol decomposition (Sobol, 1993). Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a set of n independent random variables that serve as an input to a mathematical and/or computational model $g(\cdot)$. This model produces a scalar random variable Y as output, that is, $Y = g(\mathbf{X})$, the Sobol decomposition of which is given by Sobol (1993):

$$Y = g(\mathbf{X}) = g_0 + \sum_{i=1}^{n} g_i(X_i) + \sum_{i<j} g_{i,j}(X_i, X_j) + \ldots + \sum_{i<j<k} g_{i,j,k}(X_i, X_j, X_k) + \ldots + g_{1,2,\ldots,n}(X_1, X_2, \ldots, X_n). \qquad (2)$$

In this equation, $g_0 = E[Y]$ and $g_i(X_i) = E[Y \mid X_i] - g_0$ represents the variation of Y due to the changes in the input variable X_i only when the mean g_0 has been considered. In the same way, $g_{i,j} = E[Y \mid X_i, X_j] - g_0 - g_i - g_j$ represents the variation of Y that is not accounted for by the changes in variables X_i and X_j taken separately.

Given the independence of each term (orthogonality) in the decomposition in Equation 2, the variance of the model output is equal to the sum of the contributions of variances associated with singles, pairs, triplets and so on, of input variables:

$$V[Y] = \sum_{i=1}^{n} V_i + \sum_{1 \le i < j \le n} V_{i,j} + \sum_{1 \le i < j < k \le n} V_{i,j,k} + \ldots + V_{1,2,\ldots,n}. \qquad (3)$$

From this decomposition, the variation of the output Y associated with variations in input variable X_i with no reference to other variables is given by the ratio of $V_i \,/\, V[Y]$, leading to the definition of the single effect index (Sobol, 1993) or the main factor as:

$$S1_i = \frac{V_i}{V[Y]} = \frac{Var_{X_i}[E_{\mathbf{X}_{\sim i}}[Y \mid X_i]]}{Var[Y]}. \qquad (4)$$

The variation of the output Y associated with changes when the input variables $(X_i, X_j), (X_i, X_j, X_k), \ldots$ change at the same time can be quantified with the variances $V_{i,j}, V_{i,j,k}, \ldots$ in Equation 3. Thus, the total effect index is given by Homma and Saltelli (1996):

$$ST_i = \frac{V_i + \sum_{1 \le i < j \le n} V_{i,j} + \ldots + V_{i,\ldots,n}}{V[Y]} = S1_i + \sum_{1 \le i < j \le n} S2_{ij} + \ldots + Sn_{i,\ldots,n} = 1 - \frac{Var_{\mathbf{X}_{\sim i}}[E_{X_i}[Y \mid \mathbf{X}_{\sim i}]]}{Var[Y]} \qquad (5)$$

where $\mathbf{X}_{\sim i}$ indicates fixing all input variables X_j except variable i. ST_i is a measure of the total contribution of the variable to the output including first and higher order effects (Saltelli et al., 2008).

3.3.3.1 Polynomial Chaos Expansion

The estimation of the single and total effect sensitivity indices using Equation 4 and Equation 5 can be conducted using Monte Carlo simulation, but this is computationally intensive (Sobol, 1993; Saltelli, 2002; Saltelli et al., 2008), especially for large complex socio-hydrological models. Such models require the evaluation of 2^n Monte Carlo integrals to calculate those sensitivity indices, which is only practically feasible if the number of input parameters, n, is small (Sudret, 2008). Therefore, GSA of the complex models must combine 1) an efficient approach to evaluate models, which is achieved with Hadoop-based cloud computing, and 2) an efficient approach to estimate the sensitivity indices with a metamodel or surrogate model. According to Storlie and Helton (2008) "*The use of meta-models for estimating sensitivity measures can be more accurate than the use of standard Monte Carlo methods for estimating these measures with small to moderate sample sizes*". In the GSA framework, an important type of orthogonal polynomial metamodel called Polynomial Chaos Expansion (PCE) can be used to efficiently estimate the sensitivity indices in Equation 4 and 5 (Sudret, 2008; Garcia-Cabrejo and Valocchi, 2014).

The Polynomial Chaos Expansion (PCE) is a polynomial expansion of a random variable Y in terms of other random variables ξ with a given Probability Density Function (PDF) using an orthogonal basis that depends on that PDF. According to Wiener (1938) and Ghanem and Spanos (1991), the PCE of a random variable Y can be expressed as:

$$Y(\boldsymbol{\xi}) = \sum_{j=0}^{\infty} \beta_j \Psi_j(\boldsymbol{\xi}) \tag{6}$$

where β_j are a set of coefficients that define the expansion, $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_{\mathbf{n}})$ are independent random variables with a given PDF $\mathbf{f}_{\boldsymbol{\Xi}}$, and Ψ_j are a given type of multivariate orthogonal polynomials that depends on $\mathbf{f}_{\boldsymbol{\Xi}}$. For example, if $\boldsymbol{\xi}$ follow a multivariate normal distribution then Ψ_j are the Multivariate Hermite Polynomials, while Ψ_j are Multivariate Legendre Polynomials in the case that all $\xi_i, i = 1, \ldots, n$ follow a uniform distribution. The PCE of order M and degree p are defined using these multivariate orthogonal polynomials that can be obtained by products of univariate polynomials as $\Psi_j \equiv \Psi_{\boldsymbol{\alpha}} : \Psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}) = \prod_{i=1}^{M} \Psi_{\alpha_i}(\xi_i)$, where M is usually equal to the number of random input parameters of the model (i.e., $M = n$), and α_i are the terms of an integer sequence $\boldsymbol{\alpha}$ defined as (see Sudret, 2008):

$$\boldsymbol{\alpha} = \{\alpha_i; i = 1, \ldots, M\}, \;\; \alpha_i \geq 0, \;\; \sum_{i=1}^{M} \alpha_i \leq p. \tag{7}$$

The number of terms in the PCE of Y using the multivariate orthogonal polynomial $\mathit{\Psi}_{\boldsymbol{\alpha}}$ is given by $D + 1 = (M + p)!/(M!\,p!)$. A simple example of the definition of the PCE of random variable Y using one-dimensional and multivariate orthogonal polynomials is presented in Appendix B.2. The coefficients in the PCE are estimated either by projection taking advantage of the orthogonality of the polynomials (Ghanem and Spanos, 1991; Xiu and Karniadakis, 2002 and Xiu, 2010) or by regression using a set of model evaluations (Eldred and Burkardt, 2009). Once the coefficients are estimated, the mean and variance of Y can be obtained using

$$\bar{Y} \;= \beta_0 \text{ and } \sigma_Y^2 = \sum_{j=0}^{D} \beta_j^2 \langle \mathit{\Psi}_j^2 \rangle \tag{8}$$

where $\langle \mathit{\Psi}_j^2 \rangle$ is the expected value of the square of the orthogonal polynomials used in the PCE.

3.3.3.2 GSA using PCE

The sensitivity indices can be estimated from the coefficients of the PCE of the random variable Y. The variances of Y due to changes in the input variable X_i only and the joint change of this variable and other variables $(X_i, X_j), (X_i, X_j, X_k), \ldots$ required for the estimation of the sensitivity indices $S1_i$ and ST_i can be obtained using Equation 6 after choosing the specific coefficients of the PCE of Y where X_i appears. In other words, the Sobol decomposition of Y can be obtained from a reorganization of the coefficients of its PCE (Sudret, 2008):

$$\begin{aligned} Y \;\approx \beta_{\mathrm{PC}}(\mathbf{X}) &= \sum_{j=0}^{D} \beta_j \mathit{\Psi}_j(\mathbf{X}) = \beta_0 + \sum_{i=1}^{n} \sum_{\boldsymbol{\alpha} \in \Im_i} \beta_{\boldsymbol{\alpha}} \mathit{\Psi}_{\boldsymbol{\alpha}}(X_i) + \sum_{1 \leq i_1 \leq i_2 \leq n} \sum_{\boldsymbol{\alpha} \in \Im_{i_1, i_2}} \beta_{\boldsymbol{\alpha}} \mathit{\Psi}_{\boldsymbol{\alpha}}(X_{i_1}, X_{i_2}) + \ldots \\ &+ \sum_{1 \leq i_1 < \ldots i_s \leq n} \sum_{\boldsymbol{\alpha} \in \Im_{i_1, \ldots, i_s}} \beta_{\boldsymbol{\alpha}} \mathit{\Psi}_{\boldsymbol{\alpha}}(X_{i_1}, \ldots, X_{i_s}) + \ldots + \sum_{\boldsymbol{\alpha} \in \Im_{1,2,\ldots,n}} \beta_{\boldsymbol{\alpha}} \mathit{\Psi}_{\boldsymbol{\alpha}}(X_1, \ldots, X_n) \end{aligned} \tag{9}$$

where $\Im_{(i)}$ is the multi-index of the input variable X_i and it is given by:

$$\Im_{i_1,i_2,\ldots,i_s} = \left\{ \boldsymbol{\alpha} : \begin{array}{l} \alpha_k > 0 \;\; \forall k = 1,\ldots,n,\; k \in (i_1,\ldots,i_s) \\ \\ \alpha_k = 0 \;\; \forall k = 1,\ldots,n,\; k \notin (i_1,\ldots,i_s) \end{array} \right\} \tag{10}$$

Where $\boldsymbol{\alpha}$ is an integer set defined in Equation 7 (For more details about this construction, see Sudret, 2008). Using this multi-index and Equation 8, the single effect index $S1_i$ can be compactly expressed as:

$$S1_i = \frac{V_i}{V[Y\,]} = \frac{\sum\limits_{\boldsymbol{\alpha}\in\Im_i} \beta_{\boldsymbol{\alpha}}^2 \langle \Psi_{\boldsymbol{\alpha}}^2 \rangle}{\sum\limits_{k=1}^{D} \beta_k^2 \langle \Psi_k^2 \rangle} \tag{11}$$

where the numerator is the variance of the terms of Equation 9 involving only the single variable X_i, and the denominator is the total variance of the output Y. The total sensitivity index requires the definition of a set of valid indices $\mathcal{J}_i$ that selects all the terms in the PCE where the variable X_i is present (Sudret, 2008; Alexanderian et al., 2012):

$$\mathcal{J}_{i_1,i_2,\ldots,i_s} = \{ \boldsymbol{\alpha} : \alpha_k > 0 \;\; \forall k = 1,\ldots,n,\, k \in (i_1,\ldots,i_s) \} \tag{12}$$

and the total effect index of the variable i can be compactly expressed as:

$$ST_i = \frac{V_i + \sum\limits_{1\le i<j\le n} V_{i,j} + \sum\limits_{1\le i<j<k\le n} V_{i,j,k} + \ldots}{V[Y\,]} = \frac{\sum\limits_{\boldsymbol{\alpha}\in\mathcal{J}_i} \beta_{\boldsymbol{\alpha}}^2 \langle \Psi_{\boldsymbol{\alpha}}^2 \rangle}{\sum\limits_{k=1}^{D} \beta_k^2 \langle \Psi_k^2 \rangle}. \tag{13}$$

We then carry out GSA for the coupled socio-hydrological models using PCE to measure both spatial and temporal impacts of the five target behavioral parameters, κ_{pr}, ν_{pr}, κ_{prep}, ν_{prep} and λ on model outputs, including crop profits and the groundwater table.

3.4 Results and Discussions

We assume that the five behavioral parameters are independent and uniformly distributed random variables for each agent (see Table 3-1). For example, Figure 3-5 shows the sample distribution of the behavioral parameter sets randomly selected by agent 18 (See the agent's location in Figure 3-1), and their relationships with each other. The histogram plots on the diagonal show that every behavioral parameter is uniformly distributed, and the off-diagonal plots show that there exists no correlation among each other given the scattering of the sample points. This confirms that the computational procedure we use to generate behavioral parameter sets for sensitivity analysis works properly.

We then developed two approaches to run the coupled socio-hydrological models in parallel with the Apache Hadoop framework. Approach I exploits both the independence of model evaluations and lack of explicit interactions between agents using the MapReduce framework. Table 3-2 shows the running time of the MAS model for a single year and a single execution with Approach I. The results show some improvements with different numbers of map tasks over running the model sequentially on a single machine (as shown in the "Time faster" row in Table 3-2), but not as much as we expected. Since the bulk of the execution of the coupled models happens in the execution of the MAS model, parallelizing the agents should provide a great opportunity for improving performance, but the overhead of copying the input files and starting the MapReduce tasks negates much of the performance gains that would be realized by executing the agents in parallel. If this overhead did not exist, we expect to see better performance improvements approaching 5, 10, and 20 times for 5, 10, and 20 map tasks, respectively. In other words, data locality and the overhead of initialization of MapReduce tasks are the major concerns while applying this approach to run the socio-hydrological model in parallel.

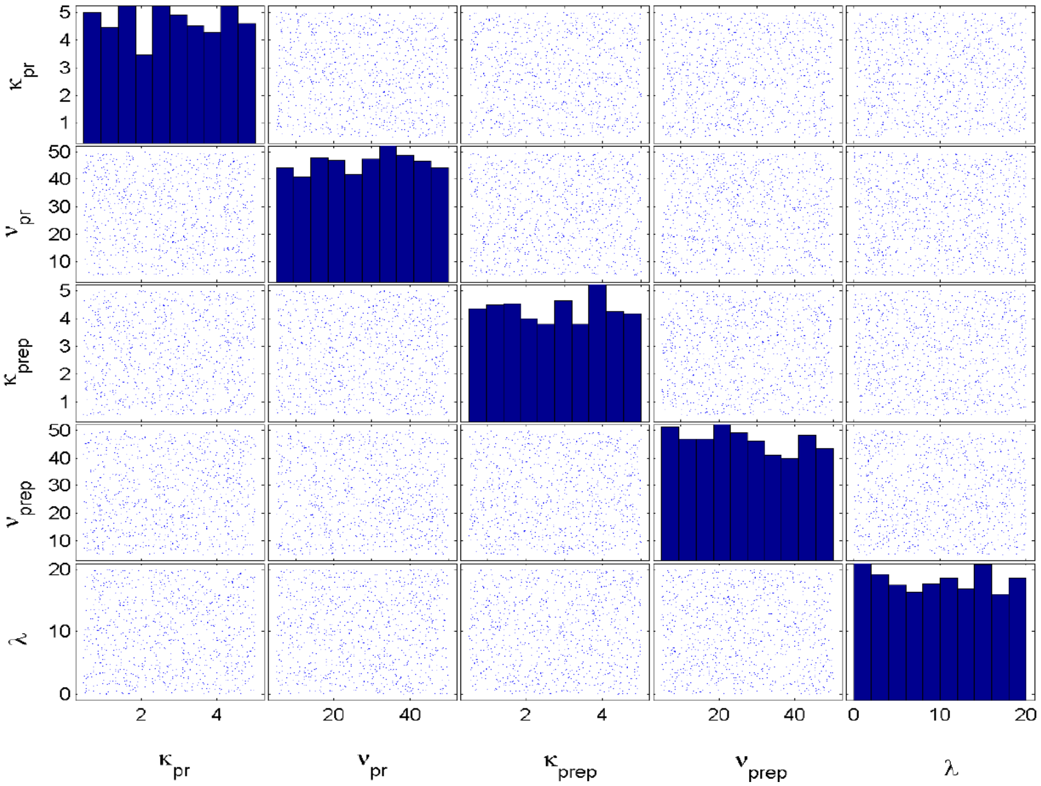


Figure 3-5 Sample distribution of the behavioral parameter (κ_{pr} , ν_{pr} , κ_{prep} , ν_{prep} and λ) and their correlation with each other. The off-diagonal blocks in the Biplot show that there exist no correlation between different behavioral parameters and the diagonal blocks show that every behavioral parameter is uniformly distributed.

Table 3-2 The execution time of the MAS model with 5, 10 and 20 map tasks for approach I.

Year	1993			1994		
Map Tasks	5	10	20	5	10	20
	79.00	64.00	62.00	85.00	75.00	76.00
Execution Time (s)	73.00	69.00	60.00	103.00	79.00	73.00
	72.00	65.00	57.00	83.00	78.00	73.00
Average	74.67	66.00	59.67	90.33	77.33	74.00
Time faster	2.25	2.55	2.82	1.86	2.17	2.27

In addition, according to the analysis based on Equation 1, for the first approach the optimal number of agents per map task is $n = (c+O)/t$. The number of map tasks, m equals to $Tt/(c+O)$ if we are not limited by the available computer resources. In our case, $n \approx 20$ and $m \approx 2,300$ with 46 agents/scenario and 1,000 scenarios. However, we can only access $m_{\text{lim}} \approx 50$ nodes on the Illinois Cloud Computing Testbed (http://cloud.cs.illinois.edu/hardware.html). As a result, this approach cannot achieve its optimal efficiency due to the limited computation resources. Different from Approach I, Approach II exploits only the independence of model evaluations using the MapReduce framework, which means that a single scenario of the coupled models is executed in its entirety over 14 years in a single map task, and no execution occurs in the reduce phase. The execution time of the MAS model for a single year with Approach II is 234.00s, much longer than the execution time with Approach I as shown in Table 3-2. However, for Approach II, first, no extra overhead to initialize a new MapReduce job for the consecutive years is required. Thus, this approach scales nearly linearly with the number of map tasks. Second, all machine nodes in the Hadoop cluster are multiple cores and the multithreaded programming technique can be used to parallelize the agents within a single map task (Hu et al., 2015). Third, the implementation of this approach is much more straightforward than Approach I. We therefore decide to use Approach II to run the coupled models for sensitivity analysis. With this approach, running 1,000 scenarios of the coupled models can be completed within two hours using the Illinois Cloud Computing Testbed, a substantial improvement over the 42 days required to run these scenarios sequentially on a desktop machine.

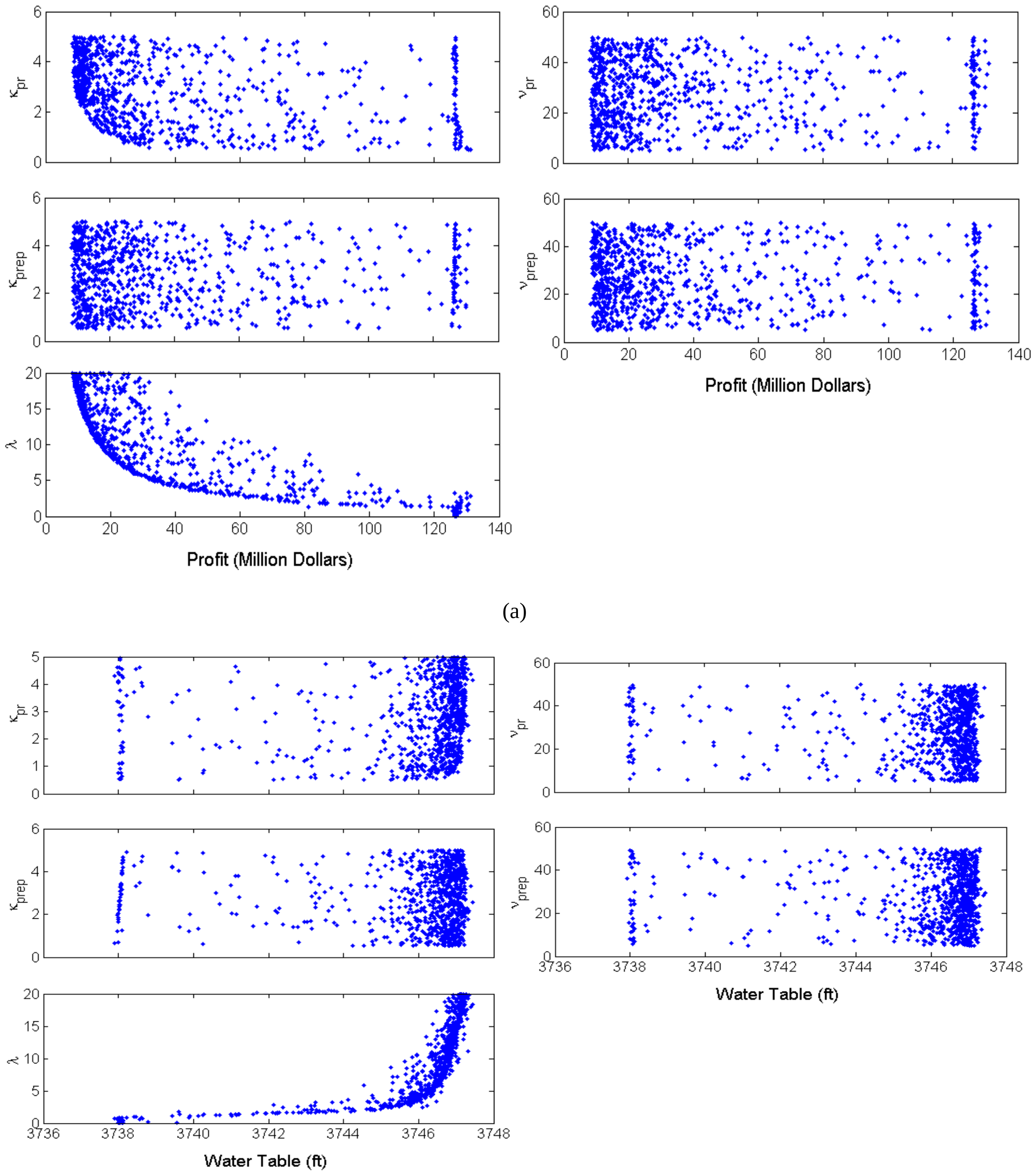
GSA using PCE requires the estimation of the PCE of the crop profits and the water table. Given the fact that the input parameters follow uniform distribution, the PCE of the output variables is defined using Legendre polynomials according to Xiu (2010). Each agent has five random behavioral parameters (i.e., $M = 5$). Given Equation 7, the sum of single-index, α_i for five random variables should be no more than the degree, p. The number of coefficients, $D+1$ in the PCE monotonically increases with p, and the number of model evaluations should be at least twice as many as the number of coefficients in the PCE for an accurate estimation of these coefficients (Eldred and Burkardt, 2009). Due to the complexity of the coupled models, we want to have the least number of model evaluations by taking the minimum value of p which satisfies

Equation 7, that is, $p = 5$. Thus, the total number of coefficients in the PCE is 252 ($M = 5, \mathrm{p} = 5$; See Sudret, 2008) and the minimum number of required model evaluations is 504. In our case, we have sufficient model inputs and outputs (as the result of 1,000 model evaluations) which can be used to derive the coefficients of the PCE-based surrogate model by the regression approach (Eldred and Burkardt, 2009) and calculate sensitivity indices of the behavioral parameters.

The behavioral parameters play important roles in agents' decisions on pumping, which are critical to the outputs of the coupled models, including crop profits and groundwater. Figure 3-6 (a) shows the relationship between the behavioral parameters and crop profits for agent 18 at a specific time in year 2006. Given the scattering points, it is found that crop profits do not change with the variation of the behavioral parameters ν_{pr}, κ_{prep} and ν_{prep}, which are also observed in the relationships between the behavioral parameters and the water table in Figure 3-6 (b). In addition, some minor impacts of parameter κ_{pr} on crop profits, but not on the water table are observed. In contrast, parameter λ has significant impacts on the variations of crop profits as well as the water table, in particular when λ is small. In other words, agents (defined with small λ values as willing to task risks) tend to have more impacts on crop profits and the water table, but more analysis needs to be done to confirm our hypothesis in the future work.

We are aware that the aforementioned statements are derived from the analysis of the relationships between the input and output variables of the coupled models at a specific time step. GSA using PCE can also help us explore the temporal evolution of the single and total effect index, $S1_i$ and ST_i of the behavioral parameters on crop profits and the water table, respectively, as shown by Figure 3-7 (a) and (b). From the values of the single effect index (continuous line in Figure 3-7 (a) and (b)), parameter λ has the dominant impact on both crop profits and the water table over the entire simulation period. The narrow gap between the continuous line and the dashed line shows that the impact of parameter λ on the coupled models comes primarily from the variation of the parameter itself, rather than the result of the interactions with other parameters. In addition, we also notice the impact of the other behavior parameters, in particular an increase in the single and total effect indices of the parameter κ_{pr} on crop profits in years 1994-1998 and 2002-2006, with a corresponding reduction in the sensitivity indices of parameter λ. As for the water table, beyond the initial noisy period (i.e. the warm-up period for the RRCA model) the variations in the period 1994-2006 are controlled mainly by the individual variation of parameter λ (single effect index as continuous purple line in Figure 3-7 (b)), and the role of the interactions associated with this variable is negligible as shown by the total effect index (dashed purple line in Figure 3-7 (b)). This result suggests that among the five behavioral variables, parameter λ (i.e., agents' attitudes towards the fluctuation of crop profits) has the greatest impact on the water table. In addition, there exists a small increase in the single and total effect indices of parameter κ_{pr} with a corresponding decrease in the importance measures of parameter λ. However, this increase is not significant enough to make κ_{pr} an influential variable. In summary, the results of the GSA applied on agent 18 indicate that the variability showed by the crop profits are controlled mainly by individual variations of parameters λ and κ_{pr}, while the variations in the water table are controlled only by the individual variations of λ.

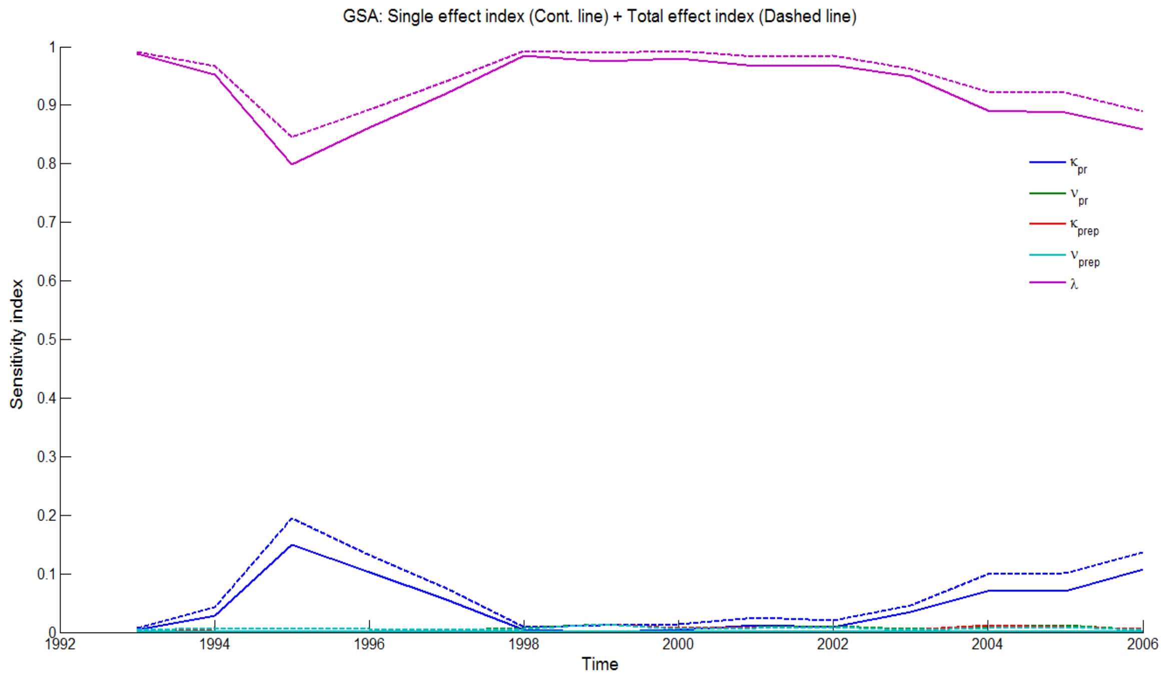
In addition, we also check the spatial evolution of the sensitivity index of the behavioral parameters for the selected agents from upstream to downstream along the Republican River (top left and right: agent 20 and 27; bottom left and right: agent 21 and 23 as shown in Figure 3-1) on crop profits and the water table. Figure 3-8 (a) shows that beyond the initial noisy warm-up period in the periods 1994-1998 and 2002-2006, the variations in the crop profits for the agents located both downstream and upstream are controlled mainly by the individual variations of parameters λ and κ_{pr}. As is evident from this plot, the variations in the crop profits due to the interactions between these parameters are small except in the case of the downstream agent where the interactions with κ_{pr} contribute much more to the variations of the crop profits than that by κ_{pr} alone. The sensitivity indices of the water table are shown in Figure 3-8 (b). The variations of the water table in the period of 1994-2002 are also controlled by the individual variations of parameter λ and κ_{pr}, and the importance of the interactions between these parameters increase downstream in such a way that for the agent located downstream the interactions account for almost 50% of the contribution to the variations in the water table. In the period 2002-2006, the influence of parameter κ_{pr} reduces with the corresponding increase in the single effect index of parameter λ and the reduction of the influence of the interactions associated with this influential parameter. These results are interesting in terms of the impacts on the variations of crop profits and the water table as the result of the dynamic interactions between parameter κ_{pr} and λ across different agents.



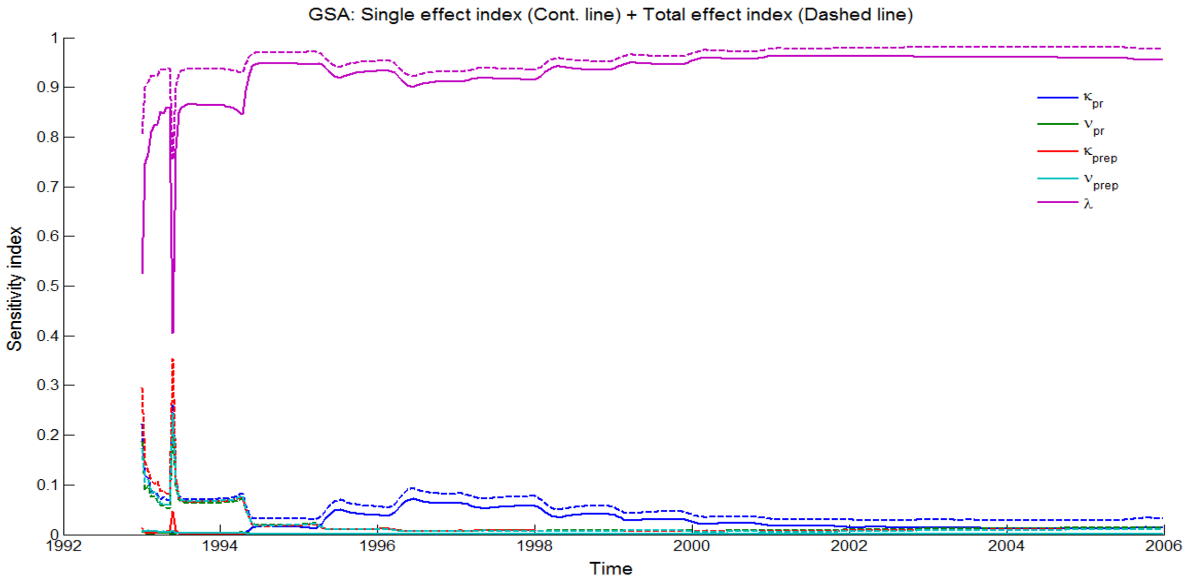
(a)

(b)

Figure 3-6 The relationship between behavioral parameters and crop profits for agent 18 in year 2006; (b) The relationship between the behavioral parameters and the water table in year 2006.

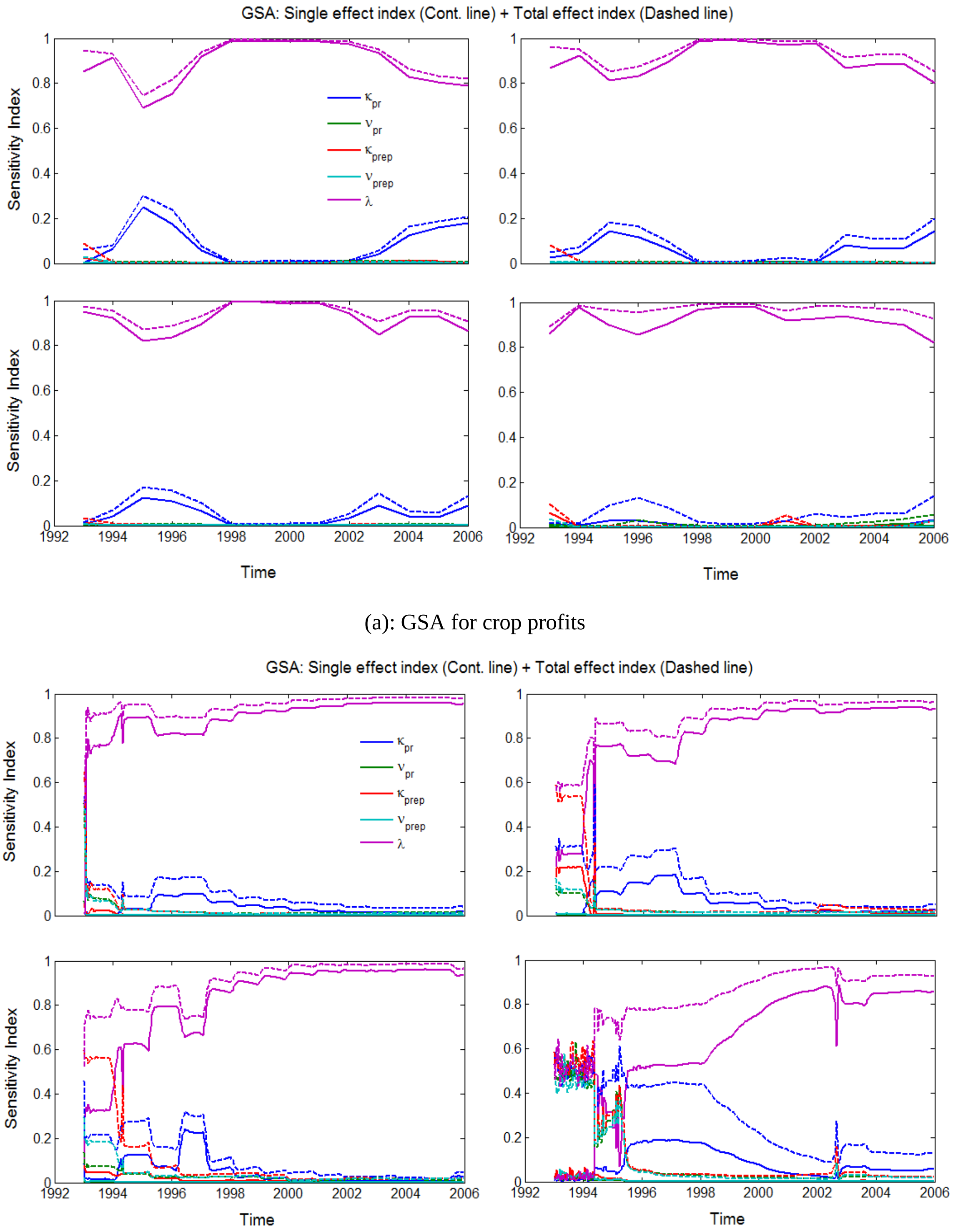


(a): GSA for crop profits



(b): GSA for water table

Figure 3-7 (a) and (b) The temporal evolution of the sensitivity index of the behavioral parameters for agent 18 on crop profits and the water table (The continuous line and dashed line denote the single and total effect index of the behavioral parameters).



(a): GSA for crop profits

(b): GSA for the water table

Figure 3-8 (a) and (b) The spatio-temporal evolution of the sensitivity index of the behavioral parameters for the selected agents from upstream to downstream of the Republican River (top left and right: agent 20 and 27; bottom left and right: agent 21 and 23 as shown in Figure 3-1) on crop profits and the water table.

3.5 Conclusions

In this chapter, a methodological framework for the application of GSA to large-scale socio-hydrological models is presented. This framework attempts to find a balance between the heavy computational burden associated with the model execution and the number of model evaluations required for GSA analysis. Specifically, the balance is achieved through the combination of Hadoop-based cloud computing and Polynomial Chaos Expansion (PCE); the former can efficiently execute a large number of complex models in parallel and the latter allows efficient estimation of sensitivity indices from PCE coefficients. To illustrate the effectiveness of the framework, we applied it to a coupled MAS decision-making model and RRCA groundwater model to investigate how the behavior parameters associated with the agents affect the outputs from the coupled models temporally and spatially, including crop profits and the water table.

Two approaches are developed to execute the coupled models in parallel using the MapReduce framework. Approach I exploits both the independence of model evaluations and lack of explicit interactions between agents using the MapReduce framework, and thus different agents in various scenarios can run their tasks simultaneously with different machine nodes. Different from Approach I, Approach II exploits only the independence of model evaluations using the MapReduce framework, and thus different scenarios (rather than different agents) are executed with different machine nodes. Through the analysis, we found that the first approach outweighs the second approach in terms of flexibility and is also more suitable for a large number of simple computational tasks with sufficient computational resources. However, the second approach is easy to implement and scale up, and can therefore be considered as a good choice for complex computational tasks with limited computational resources, as in our case study. In addition, the multithreaded programming technique can be used to take advantage of multiple cores in all machine nodes in the Hadoop cluster and parallelize the agents within a single map task for the second approach. As a result, with Approach II, a substantial reduction of the computation time is achieved, from 42 days required to run 1,000 scenarios sequentially on a desktop machine to two hours by running them on the Illinois Cloud Computing Testbed.

Each agent is defined with five behavioral parameters (i.e., κ_{pr}, ν_{pr}, κ_{prep}, ν_{prep} and λ). Parameters κ_{pr}, ν_{pr}, κ_{prep} and ν_{prep} describe the level of confidence an agent has on the prior knowledge of the mean and variance of the crop prices and precipitation, and parameter λ describes the level of aversion the agent has to risk in pursuit of higher crop profit return. These behavioral parameters affect agents' predictions of the future crop prices and precipitation through the learning process, which are used later by agents to determine the optimal pumping rates so as to maximize their utilities. With the well-represented sample sets of the behavioral parameters and the mechanism to efficiently run the coupled models 1,000 times, a large amount of crop profits and water table data are generated. The PCE is applied to generating the surrogate model for the complex coupled models using the data sets. The variance-based sensitivity indices are then calculated using the PCE coefficients. As a result, GSA using PCE-based variance decomposition approach identifies the influential parameters (i.e., κ_{pr} and λ) and quantify the spatio-temporal interactions between agents and the groundwater system through these parameters. In addition, based on the results of temporal and spatial sensitivity analysis, we will be able to narrow down our focus to these two parameters while calibrating the coupled models against the real observation data.

References

Alexanderian, A., Winokur, J., Sraj, I., Srinivasan, A., Iskandarani, M., Thacker, W. C., & Knio, O. M. (2012). Global sensitivity analysis in an ocean general circulation model: A sparse spectral projection approach. *Computational Geosciences, 16*(3), 757-778.

Axelrod, R. (1997). Advancing the art of simulation in the social sciences. *Simulating social phenomena* (pp. 21-40) Springer.

Bier, A. (2011). A.Sensitivity analysis techniques for system dynamics models of human behavior.

Borthakur, D. (2007). Hadoop distributed file system. *Apache Software Foundation,*

Carnell, R., & Carnell, M. R. (2012). Package lhs

Chattoe, E., Saam, N. J., & Möhring, M. (2000). Sensitivity analysis in the social sciences: Problems and prospects. *Tools and techniques for social science simulation* (pp. 243-273) Springer.

Dean, J., & Ghemawat, S. (2008). Mapreduce: Simplified data processing on large clusters. *Communications of the ACM, 51*(1), 107-113. doi:10.1145/1327452.1327492

Eldred, M. S. and Burkardt, J. (2009). Comparison of non-intrusive polynomial chaos and stochastic collocation methods for uncertainty quantication. *In AIAA Proceedings*, pp. 1-20.

Francos, A., Elorza, F. J., Bouraoui, F., Bidoglio, G., & Galbiati, L. (2003). Sensitivity analysis of distributed environmental simulation models: Understanding the model behaviour in hydrological studies at the catchment scale. *Reliability Engineering & System Safety,79*(2), 205-218.

Ghanem, R. G., & Spanos, P. D. (1991). *Stochastic finite elements: A spectral approach* Springer.

Garcia-Cabrejo, O., & Valocchi, A. (2014). Global sensitivity analysis for multivariate output using polynomial chaos expansion. *Reliability Engineering & System Safety, 126*, 25-36.

Happe, K. (2005). Agent-based modelling and sensitivity analysis by experimental design and metamodelling: An application to modelling regional structural change. *XIth International Congress of the European Association of Agricultural Economists.*

Hadoop Map/Reduce tutorial (2013). https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html

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

Huard, D., & Mailhot, A. (2006). A bayesian perspective on input uncertainty in model calibration: Application to hydrological model "abc". *Water Resources Research, 42*(7), W07416. doi:10.1029/2005WR004661

Hunt, R. J., Luchette, J., Schreuder, W. A., Rumbaugh, J. O., Doherty, J., Tonkin, M. J., & Rumbaugh, D. B. (2010). Using a cloud to replenish parched groundwater modeling efforts. *Ground Water, 48*(3) doi:10.1111/j.1745-6584.2010.00699.x

Hu, Y., Cai, X., & DuPont, B. (2015). Design of a web-based application of the coupled multi-agent system model and environmental model for watershed management analysis using Hadoop. *Environmental Modelling & Software,* doi:10.1016/j.envsoft.2015.04.011

Kelly, R. A., Jakeman, A. J., Barreteau, O., Borsuk, M. E., ElSawah, S., Hamilton, S. H., . . . Rizzoli, A. E. (2013). Selecting among five common modelling approaches for integrated environmental assessment and management. *Environmental Modelling & Software,47*, 159-181.

Kleijnen, J. P., Sanchez, S. M., Lucas, T. W., & Cioppa, T. M. (2003). *A user's guide to the brave new world of designing simulation experiments* Tilburg University.

Lee, P. M. (2004). *Bayesian statistics: An introduction*. Arnold Publishing, 2004. Third edition.

Liebl, F. (1995). *Simulation: Problemorientierte einführung* Oldenbourg.

McKusick, V. (2003). Final report for the special master with certificate of adoption of rrca groundwater model. *State of Kansas v.State of Nebraska and State of Colorado, in the Supreme Court of the United States, 3605*

Molle, F. (2009). River-basin planning and management: The social life of a concept. *Geoforum, 40*(3), 484-494.

Moreau, P., Viaud, V., Parnaudeau, V., Salmon-Monviola, J., & Durand, P. (2013). An approach for global sensitivity analysis of a complex environmental model to spatial inputs and parameters: A case study of an agro-hydrological model. *Environmental Modelling & Software, 47*, 74-87.

Mulligan, K. B., Brown, C., Yang, Y. E., & Ahlfeld, D. P. (2014). Assessing groundwater policy with coupled economic-groundwater hydrologic modeling. *Water Resources Research, 50*(3), 2257-2275.

Murphy, K. P. (2007). Conjugate Bayesian analysis of the Gaussian distribution. *Technical report*, University of British Columbia.

North M. J., & Macal, C.M. (2007). Managing Business Complexity: Discovering Strategic Solutions with Agent-Based Modeling and Simulation. Oxford University Press

Nielsen, M. (2009). Write your first MapReduce program in 20 minutes. Michael's main blog: http://michaelnielsen.org/blog/write-your-first-mapreduce-program-in-20-minutes/

Republican River Compact Administration (RRCA, 2003). Republican River Compact Administration Ground Water Model: http://www.republicanrivercompact.org/v12p/RRCAModelDocumentation.pdf

Pappenberger, F., Beven, K. J., Ratto, M., & Matgen, P. (2008). Multi-method global sensitivity analysis of flood inundation models. *Advances in Water Resources, 31*(1), 1-14.

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

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., . . . Tarantola, S. (2008). *Global sensitivity analysis: The primer* John Wiley & Sons.

Sobol, I. (1993). Sensitivity estimates for nonlinear mathematical models. *Math. Modeling Comput. Experiment,* 1, 407414.

Stocki, R. (2005). A method to improve design reliability using optimal latin hypercube sampling. *Computer Assisted Mechanics and Engineering Sciences, 12*(4), 393.

Storlie, C. B., & Helton, J. C. (2008). Multiple predictor smoothing methods for sensitivity analysis: Description of techniques. *Reliability Engineering & System Safety, 93*(1), 28-54.

Sudret, B. (2008). Global sensitivity analysis using polynomial chaos expansions. *Reliability Engineering & System Safety, 93*(7), 964-979.

Sweetapple, C., Fu, G., & Butler, D. (2013). Identifying key sources of uncertainty in the modelling of greenhouse gas emissions from wastewater treatment. *Water Research, 47*(13), 4652-4665.

Sweetapple, C., Fu, G., & Butler, D. (2014). Identifying sensitive sources and key control handles for the reduction of greenhouse gas emissions from wastewater treatment. *Water Research, 62*, 249-259.

Van Hemel, S. B., MacMillan, J., & Zacharias, G. L. (2008). *Behavioral modeling and simulation: From individuals to societies* National Academies Press.

Wainwright, H. M., Finsterle, S., Jung, Y., Zhou, Q., & Birkholzer, J. T. (2014). Making sense of global sensitivity analyses. *Computers & Geosciences, 65*, 84-94.

Wiener, N. (1938). The homogeneous chaos. *American Journal of Mathematics,* , 897-936.

Xiu, D. (2010). *Numerical methods for stochastic computations: A spectral method approach* Princeton University Press.

Xiu, D., & Karniadakis, G. E. (2002). The wiener--askey polynomial chaos for stochastic differential equations. *SIAM Journal on Scientific Computing*, *24*(2), 619-644.

Zhang, C., Chu, J., & Fu, G. (2013). Sobol′'s sensitivity analysis for a distributed hydrological model of yichun river basin, china. *Journal of Hydrology, 480,* 58-68.

4 CHAPTER IV

Combining human and machine intelligence to derive agents' behavioral rules for groundwater irrigation

4.1 Introduction

In the new era of water resources management, a good understanding of physical systems alone cannot guarantee the effectiveness of the policies that are drawn upon. Policy makers need to understand stakeholders' behavior to make appropriate policies that can mitigate water conflicts and promote the sustainable use of water resources. As a result, modeling their behavior, in particular their interactions with the biophysical systems, has ever been so important in the history of water resource management. Over the last decade, agents have gained in importance for the modeling of human behaviors, and agent-based models (ABMs) have been used to study the dynamics of complex systems consisting of distributed agents, and gained its popularity in both social science and economics (Arthur, 1999; Bonabeau, 2002; Tesfatsion, 2006).

The design of agent-based model follows a bottom-up, distributed approach, starting from the definition of the attributes and behaviors of individual agents, and their interactions with the surrounding environments (Ng et al., 2011; Hu et al, 2015), which allows modelers to focus on the attributes and behaviors of individuals which otherwise may not be possible using other modeling methodologies (Crooks and Heppenstall, 2012; Urban and Schmidt, 2001). Modelers can test a variety of theoretical assumptions and concepts about human behavior within the safe environment of a computer simulation (Stanilov, 2012). Thus, for coupled human and environmental systems, ABM outweighs the conventional simulation models (built based on the top-down centralized approach) in studying the system dynamics and is more likely to capture emergent phenomena arising from the interactions between human and environmental systems.

For coupled human-environment systems, the behavioral rules of agents are usually the results of combining effects of environmental, socio-economical and institutional factors. Rule-based ABMs usually assume the availability of explicit behavior rules from domain knowledge and empirical observations. Commonly used representations of expert knowledge consist of two basic forms, declarative knowledge of facts and procedural knowledge, and the latter is typically represented in IF-THEN rules (Newell 1972; Anderson 2007). Some ABM studies assume that all agents are rational and follow the general utility optimization principles (e.g., Yang et al., 2009; Ng et al. 2011). Although the ABMs based on explicit rules are meaningful for normative prediction of a system, they are often challenged by model validation with respect to the most relevant observed facts and phenomena, which is often considered as the means to justify the modelling of agents' behavior rules within ABM (Elsawah et al., 2015).

Modeling human behavior is complex. Human behavior is not random but based on our diverse knowledge and abilities, and it would not be particularly challenging if human behavior is always rational (Kennedy, 2012). However, the rationality of human behavior is affected by emotional, intuitive, or unconscious decision making processes. These processes can distort our perceptions of the environment and the likelihood of future evaluations (Loewenstein and Lerner, 2003). In this sense, the rationality of human behavior is bounded, which is the result of the limited information, limited cognitive abilities, and limited time for humans to make decisions (Simon, 1996). Regardless of the origin of the bounded rationality, agents' bounded rationality is an important source of human behavior uncertainty, which makes it difficult, if not impossible, to derive "perfect" rules for an ABM.

However, it may not be necessary to pinpoint the origin of agents' bounded rationality case by case and simulate it explicitly. Instead, this chapter proposes an alternative approach, which presents a "grey box" to simulate agents' behaviors under the influence of bounded rationality. We will later discuss how to identify the major factors relevant to the decision variable, and obtain the grey box (i.e. agents' behavioral rules) from the data sets of these factors that hold memories of agents' behavior with the data-driven approach. The grey box can then be fed by the data of the decision variable and its major factors to predict agents' decisions given these factors.

The chapter is organized for the goal to derive agents' behavioral rules considering the impact of bounded rationality using a combined data-driven approach and domain expertise. In the next section, we first present general concepts and methods that we need to use in introducing and developing our methodology. Following it, we propose a methodological framework to derive agents' behavioral rules, use a case study to demonstrate the proposed framework, and show some results. Finally, we conclude with our findings on the methodology and results.

4.2 Methodology

Agents’ behavior reflects their cognitive processes of decision-making. They may be modelled either by how decisions should be ideally made (i.e., optimization-based) or by describing how they are actually made (i.e., rule-based) (Elsawah et al., 2015). Both the optimization-based and rule-based approaches require modelers to have a good understanding of the underlying mechanism that drives agents’ decision-making, and then model the mechanism with behavioral parameters. However, these two approaches are designed to describe agents’ behavioral rules without accounting for behavioral uncertainty arising from agents’ bounded rationality. Separate techniques are usually needed for the quantification of the impacts of agents’ behavioral uncertainty, such as global sensitivity analysis (Hu et al, 2015). A holistic method from the data-driven approach perspective (e.g., statistical modeling) can be used to derive behavioral rules using both the available data and the expert knowledge while accounting for behavioral uncertainty.

Some limitations are noticed regarding the application of data-driven approaches to derive agents' behavioral rules. The first is with data availability. Although significant progress has been made in recent years to gather data for the definition of agents and the representation of their behavioral rules (Janssen and Ostrom, 2006; Robinson et al., 2007; Smajgl et al., 2011), unlike those measuring physical quantities, ways to measure human behaviors directly are limited. Some aspects such emotion and social behaviors are very difficult to measure if not unmeasurable. Conventionally, researchers use social surveys such as interviews to gather human behavioral data indirectly. Lack of sufficient data and also the quality of behavior data make derivation, validation and verification of agents’ behavioral rules difficult for ABM development (Kennedy, 2012). Furthermore, the relationships derived by a data-driven approach can be spurious due to the missing of a confounding variable, which is an extraneous variable that correlates with other variables in a statistical model. A story about the role of a confounding variable in machine learning is about “diapers and beer”: the grocery store data show that diapers and beer consumption are correlated. However, such correlation is spurious because the two variables are actually correlated with male customers, who buy beer while purchasing diapers in the same store. For complex systems, it is difficult to identify and appropriately represent all factors that affect agents' behaviors. To rule out spurious relationships, this study will incorporate the cross validation into regression tree processes and also use the expert domain knowledge.

In the following section, we will firstly introduce basic concepts and applications of a particular type of statistical models, namely probabilistic graphical models (PGMs). Then, we will delve into a specific PGM, directed information graph (DIG) and explain how it can be used to derive the causal relationships between agents' decision and the factors. Based on the DIGs for different agents, a machine learning technique called boosted regression tree (BRT) is applied to converting the DIGs to the behavioral rules for different agents.

4.2.1 Probabilistic Graphical Models

Probabilistic graphical models (PGMs) emerge as an innovative approach to organically connect different parts used to build up the complex system while ensuring the consistency of the system. PGMs are considered as the marriage between probability theory and graph theory. The probability theory side provides ways to interface models to data and the graph theory side enables humans to vividly model highly interacting sets of variables (Jordan, 1998; Koller and Friedman, 2009). They are the representations of the probabilistic relationships of the variables in a complex system (Buntine, 1996). In recent decades, there has been a large body of work on PGMs, including but not limited to, Markov networks, Bayesian networks, and factor graphs (Pearl, 1988; Koller and Friedman, 2009). For example, given the joint distribution and a specified variable ordering, the structure of Bayesian networks (i.e. directed and acyclic graph) can be found using Markov blanket properties (Pearl, 1988). However, if the variable ordering is not known, learning and optimally approximating the structure becomes NP-hard (Chickering et al., 1994). In addition, some researches are focused on identifying causal relationships using Bayesian networks (Koller and Friedman, 2009, Ch. 21), which requires the use of expert domain knowledge to label the variables. Thus, the resulting Bayesian network depends on the variable labeling; without expert labeling the Bayesian network is not unique and the identified relationships are only correlative. For the setting of time-series, dynamic Bayesian networks can be applied to finding a Bayesian network to characterize relationships over time. Each process corresponds to multiple nodes in the graph, one for each time step (Koller and Friedman, 2009, Ch. 6). Thus, the number of potential edges increases quickly with the number of time-series, making structure learning challenging. In this chapter, we will later discuss a more recent class of PGMs where each time series is a single node with edges corresponding to causation between time series, not correlation between variables like other PGMs.

PGMs are widely used in various fields including, but not limited to, medical diagnosis, navigation, image processing and communication. Recently, a few case studies have been conducted in land and watershed management in the context of adaptive natural resource management using PGMs (Alexandridis 2006; Carmona, et al., 2011). For example, Aalders (2008) tries to incorporate the characteristics of land managers with Belief Networks (BNs) to explore the impacts of their behaviors in decision-making processes. However, they usually obtain the structure of the graphical models purely based on the domain expertise. In the following, we will propose a methodological framework that combines the expert domain knowledge with a PGM-based data-driven approach to derive the causal network structure of factors that are likely to affect agents' behavior. The derived graph structure will be translated into agents' behavioral rules for the design of ABM, which can be coupled with environmental models to investigate the interactions between human and environmental systems.

4.2.2 Granger Causality

We now discuss the framework leading to the graphical model we will use in this work. Granger proposed the definition of causality for a network of autoregressive time series in the 1960's: "Given a pair of random processes **X** and **Y**, we say that **X** is causing **Y** if we are better able to predict [the future of] **Y**, using all available information than if the information apart from [the past of] **X** had been used" (Granger 1969). The key principle is that if "**X** causes **Y**," then the past of **X** should help predict the future of **Y**. This was based on earlier time-series prediction work by Wiener (1956). Granger suggested using the ratio of model error variances as a strength of causality, the "Granger causality test". This is a statistical hypothesis test for linear models. However, for coupled human-environmental systems, there exist complex, non-linear relationships between the factors the conventional Granger causality test approach cannot deal with. In this sense, the information theoretic quantity known as directed information will be able to capture such non-linear relationships to determine Granger causality.

4.2.3 Directed Information

Given a pair of random processes **X** and **Y**, the directed information (DI) from **X** to **Y** is defined as

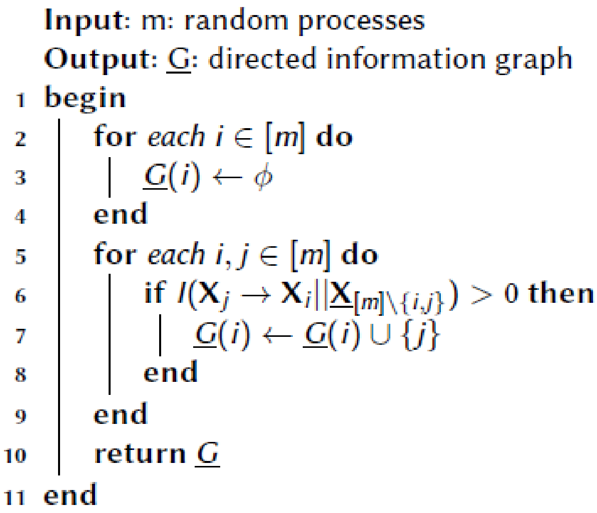
$$\mathbf{I}(\mathbf{X} \to \mathbf{Y}) := \sum_{\mathbf{t}=1}^{\mathbf{n}} (\mathbf{X}^{\mathbf{t}\text{-}1}; \mathbf{Y}_{\mathbf{t}} \mid \mathbf{Y}^{\mathbf{t}\text{-}1}) := \sum_{\mathbf{t}=1}^{\mathbf{n}} \mathbf{E}_{\mathbf{P}_{\mathbf{X}^{\mathbf{t}\text{-}1},\mathbf{Y}^{\mathbf{t}}}} \left[\log \frac{\mathbf{P}_{\mathbf{Y}_{\mathbf{t}}|\mathbf{Y}^{\mathbf{t}\text{-}1},\mathbf{X}^{\mathbf{t}\text{-}1}}}{\mathbf{P}_{\mathbf{Y}_{\mathbf{t}}|\mathbf{Y}^{\mathbf{t}\text{-}1}}} \right], \tag{1}$$

which measures how correlated the past of **X** is to the future of **Y** in terms of *Kullback-Leibler* (KL) divergence (Marko, 1973; See Appendix C.1). Different from mutual information which quantifies correlation (See Appendix C.1), DI is used to quantify the statistical causation between factors in the sense of Granger causality, even for factors that have complex, non-linear relationships. Equation 1 is well defined for any distribution **P**. However, even when measured with DI, spurious relationships as the result of latent confounding factors cannot be avoided. Thus, when necessary, expert domain knowledge (i.e. human intelligence) is applied to rule out impossible directions of influence. Also, to avoid overfitting, we use a model complexity penalty known as minimum description length (MDL; Grunwald, 2007). This penalty ensures the influences found are not due to noisy data.

4.2.4 Directed Information Graph

We next discuss a PGM defined using directed information. For a set of random processes $\underline{\mathbf{X}}$, the directed information graph (DIG) is a directed graph where each node represents a process and there is a directed edge from process $\mathbf{X}_{\mathbf{j}}$ to X_{i} (for $\mathbf{i}, \mathrm{j} \in \{1, \ldots, \mathbf{m}\}$) if $\mathbf{I}(\mathbf{X}_{\mathbf{j}} \to \mathbf{X}_{\mathbf{i}} \| \underline{\mathbf{X}}_{\{1,\ldots,\mathbf{m}\}\backslash\{\mathbf{i},\mathrm{j}\}}) > 0$ (Quinn et al., 2011; Amblard and Michel, 2011); that is to say, there is a directed edge from $\mathbf{X}_{\mathbf{j}}$ to $\mathbf{X}_{\mathbf{i}}$, if and only if knowing the future of $\mathbf{X}_{\mathbf{i}}$ depends on the past value of $\mathbf{X}_{\mathbf{j}}$, even when conditioned on the past of all other processes in the network. We now use DIG to derive the causal relationship of variables that are likely to affect agents' behavior.

We define a directed information graph, $\mathbf{G} := (\mathbf{V}, \mathbf{E})$ where the set of vertices is denoted by **V** and the set of edges is denoted by **E**. Each process is represented as a vertex and the DIG, **G** is identified by testing the direction information from process $\mathbf{X}_{\mathbf{j}}$ to $\mathbf{X}_{\mathbf{i}}$ conditioned on the past of all other processes in the network. We calculate the error variances of model fitting under two scenarios: I) $\mathbf{X_i}$ is the response variable and the rest of the vertices in the network are explanatory variables; II) $\mathbf{X_i}$ is the response variable and the rest of the vertices except $\mathbf{X_j}$ in the network are explanatory variables. The corresponding error variances for scenario I and II are denoted by $\sigma_{\mathbf{I}}$ and $\sigma_{\mathbf{II}}$. We then compare the logarithm of the ratio of error variances $\sigma_{\mathbf{I}}$ to $\sigma_{\mathbf{II}}$ with the minimum description length (MDL) to determine if $\mathbf{I}(\mathbf{X_j} \rightarrow \mathbf{X_i} \| \underline{\mathbf{X}}_{\{1,\ldots,\mathbf{m}\}\backslash\{\mathbf{i},j\}}) > 0$ and repeat the procedure for all the vertices in the network. As a result, we obtain the DIG as described by Algorithm I from Quinn et al. (2015):



```
Input: m: random processes
Output: $\underline{G}$: directed information graph
begin
    for each $i \in [m]$ do
        $\underline{G}(i) \leftarrow \phi$
    end
    for each $i, j \in [m]$ do
        if $I(\mathbf{X}_j \rightarrow \mathbf{X}_i || \underline{\mathbf{X}}_{[m]\backslash\{i,j\}}) > 0$ then
            $\underline{G}(i) \leftarrow \underline{G}(i) \cup \{j\}$
        end
    end
    return $\underline{G}$
end
```

Algorithm I: construction of the directed information graph.

The computational complexity of Algorithm I is **O**(m^2) if the directed information values are calculated beforehand. The algorithm can also be implemented in parallel, since each DI values can be computed separately. In order to rule out the impossible directions of influence, we will integrate the expert domain knowledge into Algorithm I. Once the DIG, **G** is derived, we then select the target vertices used to describe agents' behavior, keep the incoming edges of these target vertices and their connected vertices, and prune the rest of the edges and vertices. In this way, we obtain causal relationships between the selected vertices and target vertices in terms of simplified DIG. These causal relationships are later used to define the behavioral rules of agents.

Based on the similarity of the simplified DIGs for various agents, a graph clustering tool, GraphCluster is used to cluster the graphs of individual agents based on existing similarities (Reforgiato et al., 2008). The clustering algorithm proceeds in two phases: 1) finds the highly connected substructures (i.e., the shortest path from one vertex to the other vertices in the graph) in each graph; 2) uses those substructures to represent each graph as a feature vector. Clustering itself is done using the *k*-means method (Lloyd, 1982). As a result, we can use fewer graphs to represent agents' behavioral rules and reduce the computation for the agent-based modeling.

4.2.5 Boosted Regression Tree

Boosted Regression Tree (BRT) is a machine learning technique that uses the boosting techniques to combine a large number of relatively simple tree models adaptively to optimize the predictive power of the ensemble tree models (Roe et al., 2005, Elith et al., 2008 and Pedregosa et al., 2011). Different from conventional regression methods that aim to produce a single best predictive model, BRT uses the boosting method to find and average many models to improve the model accuracy (Elith et al., 2008). For example, given a regression problem, at each step, boosting adds a new model to improve the predicative performance of the current models, which is measured by the deviance between the sample data and the fitted values, namely loss function. The final BRT model is the linear combination of many tree models. Several software packages have implemented BRT with easy-to-use functions (e.g., R and MATLAB). In order to use these functions, users only need to provide two important parameters: 1) the learning rate (lr) which determines the contributions of each tree model to the ensemble models; 2) the tree complexity (tr; e.g., tree depth) which controls whether interactions are fitted (Elith et al., 2008).

Given the empirical comparison of supervised learning algorithms done by Caruana and Niculescu-Mizil (2006), the BRT model has overall the best predictive performance over the other methods. Thus, BRT is chosen to convert the DIGs for different agents that describe the causal relationships between the selected vertices and target vertices, to ensemble tree models. As a result, these ensemble tree models are then used as grey boxes to represent agents' behavioral rules, and simulate/predict agents' decisions as shown by Equation 2:

$$\mathbf{D} \equiv \mathbf{BRT}(\underline{\mathrm{V}}) + \delta, \tag{2}$$

where **D** represents the decision variable and $\underline{\mathrm{V}}$ is a set of variables identified by the directed information graph algorithm and used as the input to the ensemble tree models, **BRT**. The residual δ reflects the information contributed to the decision variable by the variables that are missed out, such as, constant variables like soil types.

4.3 Case Study Site

The Republican River originates in the high plains of northeastern Colorado, western Kansas and southern Nebraska, which covers approximately 25,018 square miles (~16 million acres) of the three states, and is encompassed by the underlying High Plains aquifer hydrological observatory (HO) area. Intensive agriculture development in the Republican River Basin since the 1970s has led to the significant increase of groundwater use for irrigation. Water conflicts and lawsuits have arisen as the result of sharing the groundwater resource among these three states. As part of the US Supreme Court settlement, a comprehensive groundwater model, the Republican River Compact Administration (RRCA) groundwater model was developed through the collaboration of the three affected states, the U.S. Geological Survey, and the U.S. Bureau of Reclamation (McKusick, 2003). MODFLOW-2000 with additional modules, a finite difference groundwater flow simulation code is used to construct the RRCA model (Harbaugh, 2005).

In our case study, a human behavioral model (i.e., ABM) is designed and coupled with this physically-based environmental model for groundwater simulation (i.e., RRCA model). The ABM provide decisions on the annual groundwater pumping rates, which are then used as inputs to the RRCA model. The coupled ABM and RRCA model are used to investigate the impacts of farmers' pumping decisions on the groundwater systems, and the simulation period is from year 1993 to 2006, as shown by Figure 4-1. Due to the fact that most data (e.g. crop areas and production) is only available at county level, a county located in the High Plains aquifer HO area is thus defined as a super-farm agent as shown by Figure 4-2, leading to 46 agents in total.

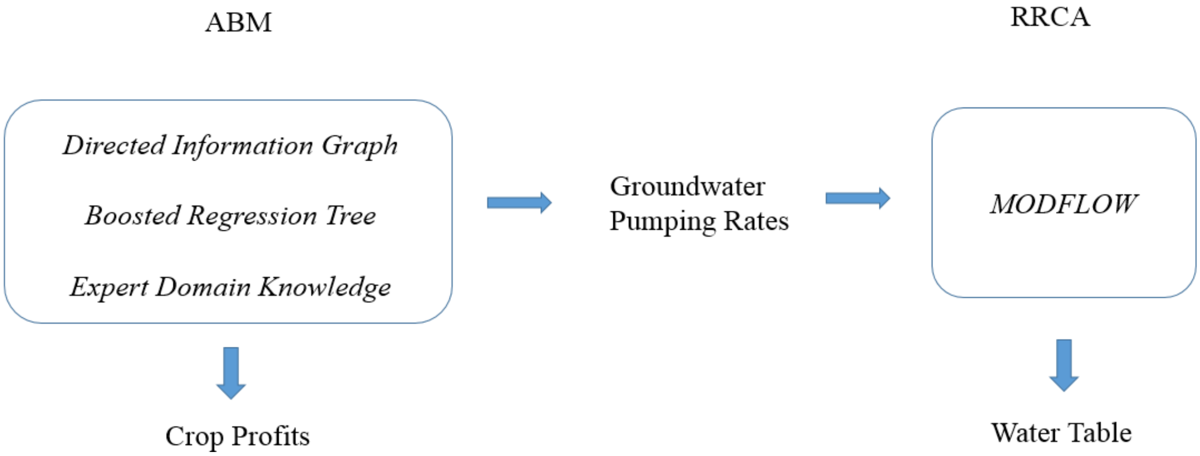


Figure 4-1 Coupling of the agent-based model (ABM) with the RRCA groundwater model. The ABM is developed through the combination of the data-driven approach (i.e., DIG and BRT) with expert domain knowledge, and the RRCA model is simulated by MODFLOW-2000. The simulation period is from year 1993 to 2006.

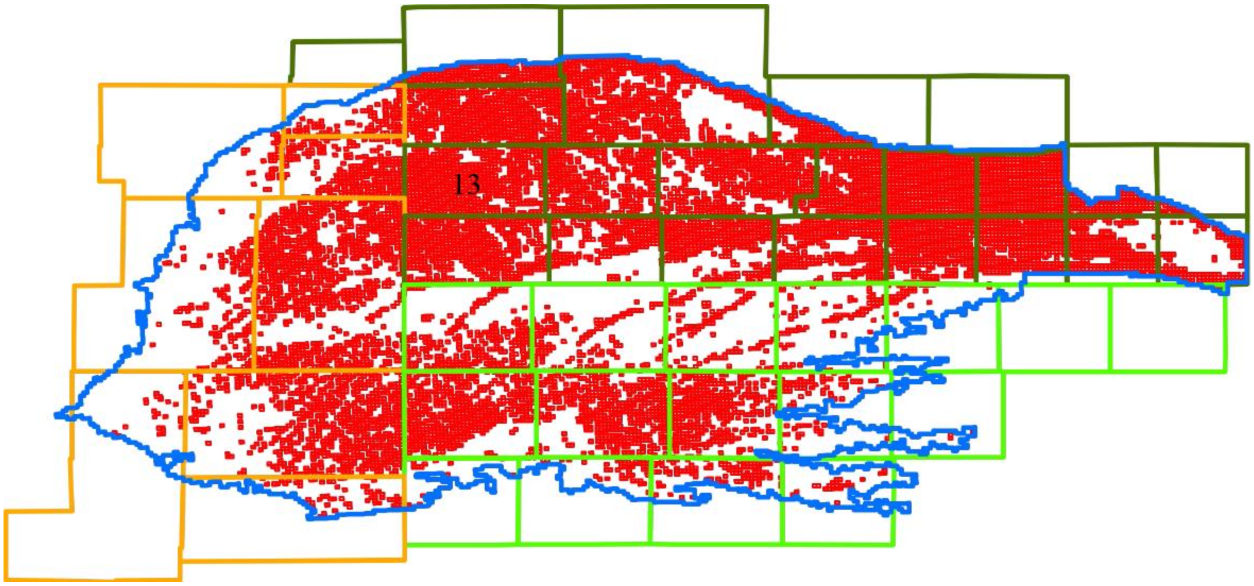


Figure 4-2 The plain view of the pumping wells (red dots) and High Plains aquifer (blue line) in MODFLOW-2000 and the overlapping counties (blocks) of different states (orange: Colorado; light green: Kansas; spruce green: Nebraska; each county is treated as an agent and the number is the selected agent ID; see Hu et al., 2015).

Table 4-1 shows the environmental, economic, social, and infrastructure factors on agents' decision on groundwater pumping depth. Note that corn price instead of other crop prices is selected as one factor in the economic category due to the fact that corn is the dominant crops in the study area (See Figure 4-7 (e)). Most of the data used by the ABM are publicly accessible from the RRCA website (http://www.republicanrivercompact.org); Carbon Dioxide Information

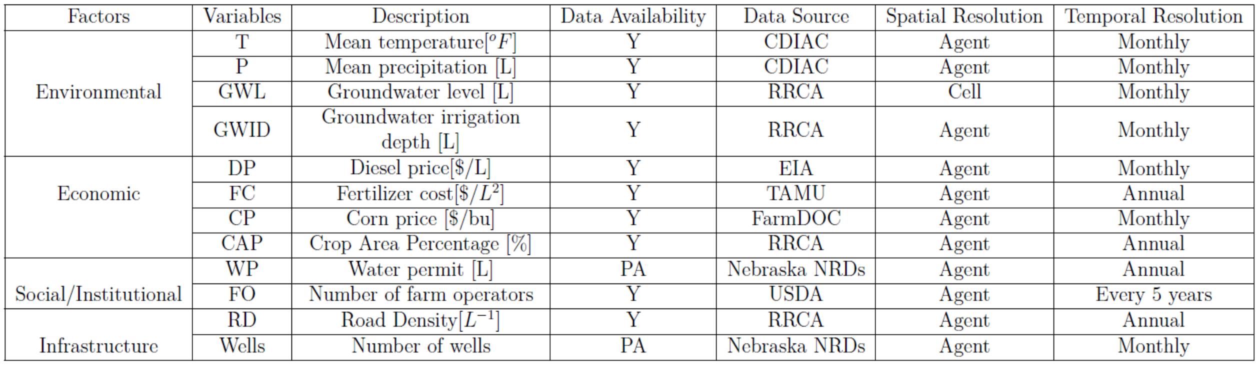
Analysis Center (CDIAC); Texas A&M AgriLife Extension (TAMU); U.S. Department of Agriculture (USDA); Farm Decision Outreach Central (FarmDOC) at the University of Illinois and Natural Resources Districts, Nebraska, U.S. We select these time-series variables based on three criteria: 1) data is available at the desired spatial and temporal scale (e.g., annual or monthly data for each agent); 2) there exist some variations in the data for each variable, since the statistical causality between variables is identified based on the variables with variations (the variables which do not change significantly over time are reflected in the form of residuals see Equation 2); 3) the matrix consisting of the selected variables (i.e., each variable is treated as a column) is full column rank. In this sense, no variables can be directly calculated from the other variables using physically-based equations or models.

We select an agent in the High Plains aquifer HO area to illustrate the application of the combined DIG and the expert domain knowledge to derive agents' behavioral rules. For example, Chase County in the Nebraska portion (i.e., agent 13 as shown by Figure 4-2) is a good candidate due to the heavy monitoring in that area. In order to account for the interactions between the Chase County agent and the neighboring agents, we also include the groundwater levels (GWLs*) of the neighboring agents as the variables to test if they have potential effects on Chase County agent's behavior. We then apply the DIG algorithm to deriving the DIG of those variables that are associated with the agents' behavior.

Next, we select the target variable among the associated variables that describes agents' pumping behavior, that is, the monthly groundwater irrigation depth (GWID). Based on the DIG obtained from the previous step, we keep the incoming edges of GWID and the associated nodes, and truncate the remaining edges and nodes. To do so, we obtain a directed acyclic graph which characterizes the causal relationships between the selected variables and GWID. This directed acyclic graph that models the agent's behavioral rules is used to simulate the agent's decision on irrigation depth. We repeat the aforementioned procedures for the other agents to obtain directed acyclic graphs to simulate their pumping behavior. With a directed acyclic graph for each of the agents, we exam the similarities among the graphs using GraphCluster. As a result, some representative graphs are identified to represent the behavior rules for different clusters of agents. Next, given the representative graph structure and the information of the associated nodes for an individual agent, we can then compute GWID using BRT. Note that the tree models are trained individually using available data sets of the nodes of the representative graph for a specific agent.

Thus, they vary from one agent to the other and forms the agent-based model, which is then coupled with the RRCA model to investigate the impacts of agents' pumping behavior on the groundwater system.

Table 4-1 List of variables associated with agent's pumping behavior.



Factors	Variables	Description	Data Availability	Data Source	Spatial Resolution	Temporal Resolution
Environmental	T	Mean temperature[^oF]	Y	CDIAC	Agent	Monthly
	P	Mean precipitation [L]	Y	CDIAC	Agent	Monthly
	GWL	Groundwater level [L]	Y	RRCA	Cell	Monthly
	GWID	Groundwater irrigation depth [L]	Y	RRCA	Agent	Monthly
Economic	DP	Diesel price[\$/L]	Y	EIA	Agent	Monthly
	FC	Fertilizer cost[\$/L^2]	Y	TAMU	Agent	Annual
	CP	Corn price [\$/bu]	Y	FarmDOC	Agent	Monthly
	CAP	Crop Area Percentage [%]	Y	RRCA	Agent	Annual
Social/Institutional	WP	Water permit [L]	PA	Nebraska NRDs	Agent	Annual
	FO	Number of farm operators	Y	USDA	Agent	Every 5 years
Infrastructure	RD	Road Density[L^{-1}]	Y	RRCA	Agent	Annual
	Wells	Number of wells	PA	Nebraska NRDs	Agent	Monthly

Y: Yes; N: No; PA: Partially Available; NA: Not Applicable; RRCA: Republican River Compact Administration; CDIAC: Carbon Dioxide Information Analysis Center; EIA: U.S. Energy Information Administration; TAMU: Texas A&M AgriLife Extension; USDA: U.S. Department of Agriculture; FarmDoc: Farm Decision Outreach Central; Nebraska NRD: Natural Resources Districts, Nebraska; Cell: 1 mile by 1 mile. There are 13,220 grid cells in total which locate in 46 different agents. The number of the cells in the individual agent varies from few to approximately 1,000.

4.4 Results and Discussion

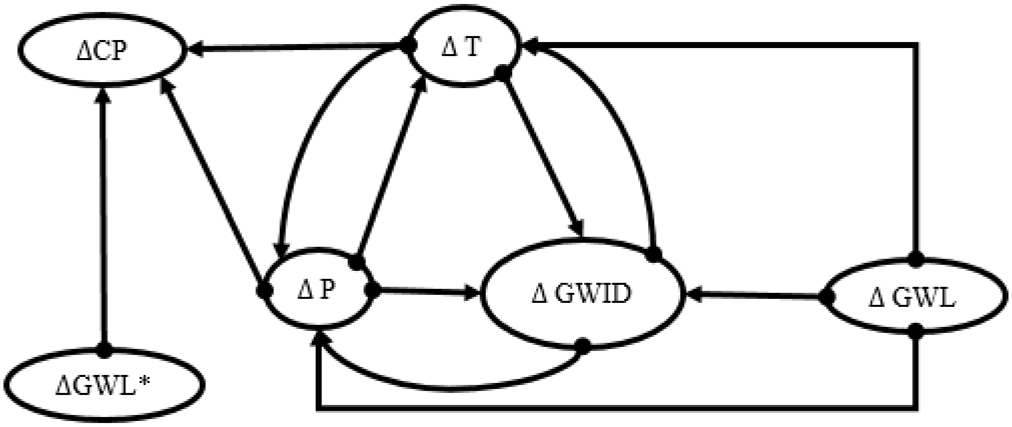


Figure 4-3 Directed information graph of variables for agent 13 in Nebraska as shown in Figure 4-2; the symbol Δ indicates the causal relationship is identified based on the variables with variations.

Without taking the expert domain knowledge into account, Figure 4-3 shows the DIG of the variables in Table 4-1 using Algorithm I. For some cases, it is intuitive for domain expert to understand the causal relationship between variables. For example, monthly mean temperature (T) and precipitation (P) causally influence the monthly groundwater irrigation depth (GWID), since temperature and precipitation can affect crop evapotranspiration (ET) and effective rainfall (ER), which is defined as the part of the rainfall stored in the root zone and can be used by crops. If the contribution of ER to crop water demand increases, the amount of water needed for irrigation (i.e., GWID) decreases accordingly.

Meanwhile, some causal relationships discovered by the algorithm are not straightforward to understand. For example, agents' irrigation behavior can affect the key components of regional climate, such as evapotranspiration, temperature and precipitation. Figure 4-3 shows that GWID causally influence the regional T and P at agent 13, which can be explained that soil moisture can increase dramatically during the warm season due to heavy irrigation. High soil moisture level can then lead to the increases in ET, cooling of surface temperature and enhancement of precipitation (Eltahir and Bras, 1996; Eltahir, 1998; Vörösmarty and Sahagian, 2000; Pielke, 2001; Kanamitsu and Mo, 2003; Betts, 2004; Haddeland et al., 2006a; Kustu et al., 2010). Similar results found by Chase et al. (1999), who investigated the effect of land use changes on the regional climate of northern Colorado plains, show that irrigational practices can introduce the forcing strong enough to affect the regional temperature, cloud cover, precipitation and surface hydrology. These scientific findings help verify some parts of the graph derived completely from the historic data using the DIG algorithm.

Figure 4-3 also shows some spurious causal relationships derived purely from data that do not make sense according to the expert knowledge. For example, T, P and GWLs* can causally influence corn price (CP), and GWL can affect the regional mean T and P. As mentioned above, these spurious relationships are ruled out either by cross-validation or the negation normal form of expert knowledge. For example, local T should not causally influence national CP, which should not be causally affected by GWLs* as well. In the case study, we try to include as few constraints as possible and do not impose any causal relationships directly based on the expert knowledge but not reflected in the data. As a result, Figure 4-4 shows the DIG of the variables that affect the agents' behavior based both on the improved BRT models with cross-validation and the expert domain knowledge.

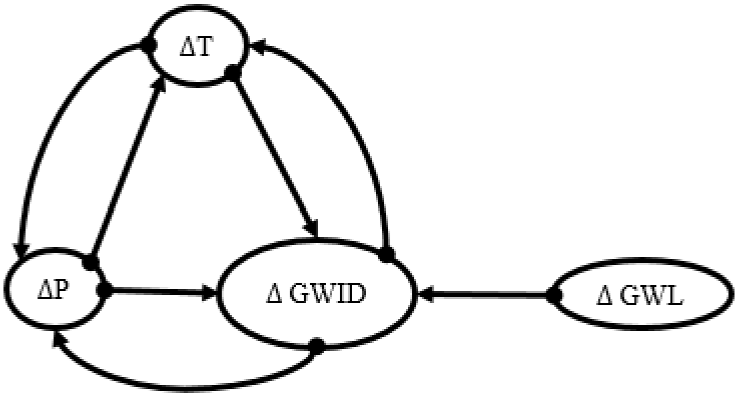


Figure 4-4 Directed information graph of the variables that affect agents' behavior based on the combination of the improved BRT models with cross-validation and the expert domain knowledge.

Given the goal to derive the DIG that shows the causal relationships between the selected variables and the target variable (i.e., GWID), based on Figure 4-4, we keep the incoming edges of GWID and the associated nodes (i.e., GWL, T and P), and truncate the rest edges and nodes. Figure 4-5 shows the nodes that have the direct influence on GWID for the Chase County agent. The value on each edge is the measure of directed influence from one variable to the other. The large the value, the more influence has one variable on the other variable. Thus, T has the dominant impact on GWID for the Chase County agent. We further generate directed acyclic graphs for all agents within the High Plains aquifer HO area, and cluster the graphs based on their similarity. As a result, three representative graphs are identified to represent all agents' decision on GWID as shown in Figure 4-6. Four factors including CP, GWL, T and P have causal influences on agents' decision on GWID to various extents, and T is the most common factor which appears in all three graphs. Note that different colors representing different agents' behavior rules are not randomly distributed, rather than display certain types of spatial patterns. In the following, we attempt to explain the formation of the spatial patterns.

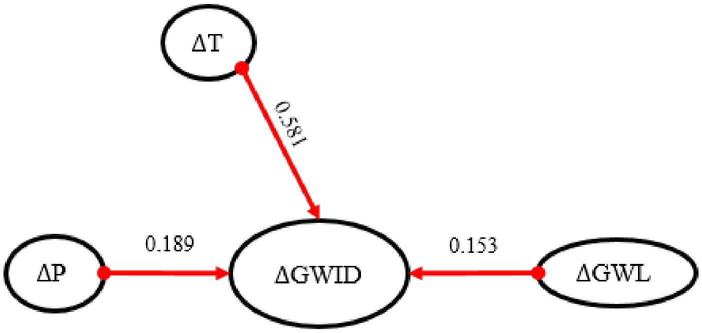


Figure 4-5 Directed information graph of Chase County Agent's decision on the groundwater irrigation depth. The numbers on the edges are the measures of directed influence from one variable to the other.

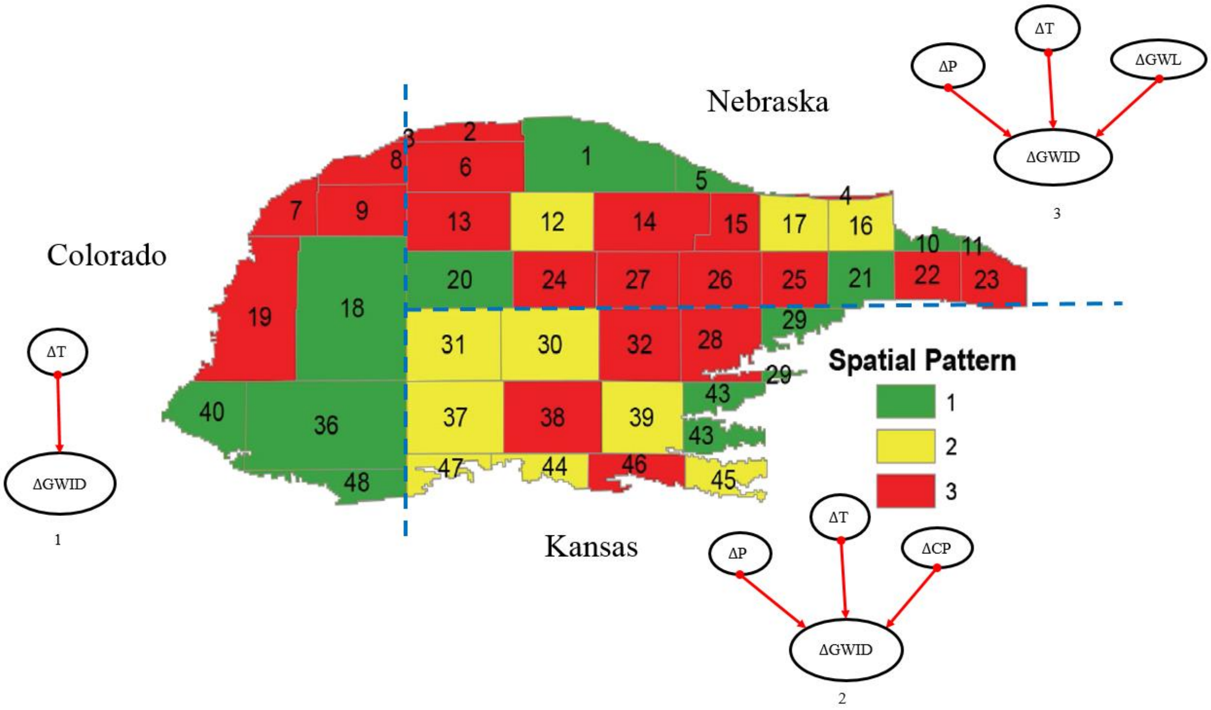


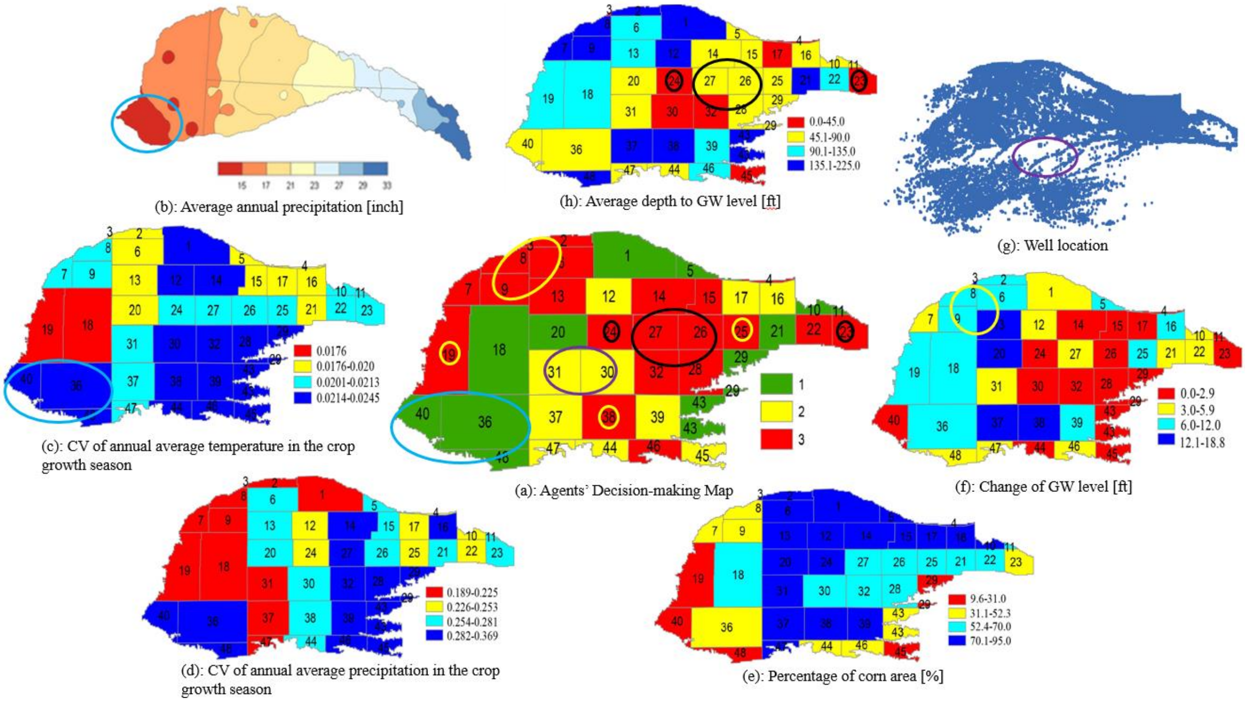
Figure 4-6 Color-coded map with three directed acyclic graphs that represent agents' decisions on the groundwater irrigation depth within the High plains aquifer HO area. The dashed lines are boundaries between different states.

The three representative DIGs over the study site is explained with some spatially distributed factors as shown in Figure 4-7 (b-h). For the agents with type 1 DIG (marked in green), T is the only factor that causally influences agents' decision on GWID. The areas of agents 36, 40 and 48 circled in blue in Figure 4-7 (a) overlap with the region circled also in blue in Figure 4-7 (b), which receives the least average annual precipitation in the Republican River basin. This can be explained as follows: in a dry area, the crop evapotranspiration (ET) which determines the crop water requirement is mainly affected by temperature. In this sense, an agent's response to temperature leads water application to satisfy crop water requirement. In addition, these three agents are among the agents with the largest coefficient of variation (CV) of annual mean temperature as shown in Figure 4-7 (c). Different from these three agents, agents 1 and 18 have relatively high but stable annual precipitation. As shown in Figure 4-7 (d), their CVs of annual mean precipitation in the crop growth season are far below the average. Thus, the small CVs in precipitation lead famers' attention more to temperature in their pumping decisions.

With type 2 DIG, T, P and CP are the factors that causally influence agents' decision on GWID. All agents with this type are shown in yellow in Figure 4-6. For most of these agents (except agents 30, 44 and 45), their corn acreage is over 70% of their cropland area as shown in Figure 4-7 (e). It explains why these agents' pumping decisions are sensitive to the variations of CP. In addition, for most of the agents (except for agents 16 and 37), they experience minor groundwater drawdowns as shown in Figure 4-7 (f); in particular, agents 30 and 31 (circled in purple) in Figure 4-7 (a) have the lowest well density over the study area as shown in Figure 4-7 (g). This explains why GWL has limited effects on these agents' pumping behavior.

For type 3 where T, P and GWL are the factors that causally influence agents' decision on GWID (agents marked in red in Figure 4-6). Referring to Figure 4-7 (f), in areas of agents 2, 3, 6, 8, 9, 13, 19, 25 and 38 with type 3 DIG (circled in yellow), the change of groundwater level is relative large; Different from these agents, agents 14, 15, 23, 24, 26, 27, 28 and 32 (circled in black) are located within the areas with small depth to groundwater level as shown in Figure 4-7 (h) and shallow pumping wells are used in these areas. Thus, although there are relative small groundwater drawdowns, they can still be noticeable to these agents with shallow pumping wells. Thus, for all the aforementioned agents their pumping decisions are affected by the variations of GWL.

We are aware of the limitations in explaining the spatial patterns of the DIGs, although we can provide some justifications with some patterns shown in Figure 4-7. Agents' decision-making is very complicated and the three simple DIGs may not represent all agents' behaviors perfectly. In addition, the data for analysis can be noisy or the sampling frequency can be insufficient or even, the important variables associated with agents' decision can be missing. Thus, it is reasonable that for some agents like agents 5, 10, 11 and 20, their DIGs are not well explained by our hypotheses.



(b): Average annual precipitation [inch]

(h): Average depth to GW level [ft]

(g): Well location

(c): CV of annual average temperature in the crop growth season

(a): Agents' Decision-making Map

(f): Change of GW level [ft]

(d): CV of annual average precipitation in the crop growth season

(e): Percentage of corn area [%]

Figure 4-7 Relationships between agents' behavior rules represented in different colors in (a) and the other graphs, (b)-(h); (b): average annual precipitation for the Republican River basin [inch]; (c): coefficient of variation (CV) of annual average temperature in the crop growth season; (d): coefficient of variation (CV) of annual average precipitation in the crop growth season; (e): distribution of the percentage of corn acreage [%]; (f): change of groundwater level [ft]; (g): locations of pumping wells in blue dots in MODFLOW; (h): average depth to groundwater level [ft]; The simulation period is from year 1993 to 2006.

The representative graphs are converted to three types of ensemble tree models using BRT. Agents' monthly GWID is then computed through these tree models. Through trial and error, we set the number of trees equal to 500, the learning rate equal to 0.01 and the tree depth equal to 4. 60% of data are used for model training and the rest data for model validation. Figure 4-8 shows the tree No. 1 of 500 trees for agent 24 and the ensemble tree model **M** is the linear combination of the 500 trees. Being different from T and P, the variation on GWL is hardly noticeable to agents if the pumping wells are not drying out. Thus, before they notice groundwater drawdowns in pumping wells, the more groundwater agents withdraw, the lower is the GWL. Thus, lower GWL corresponds to the higher groundwater withdrawal (i.e., GWID), as shown in Figure 4-8.

Given the tree model **M**, we can then compute GWID. Figure 4-9 shows the comparisons of monthly GWID between the observation and the model simulation using the ensemble tree models of agents 17, 18 and 24 between year 1993~2006. Notice that we only consider GWID during the crop planting/growing season from May to October, that is, 84 months in total for 14 years. It shows that the results for validation from the ensemble tree models have a good match with the observation one in general, although the spike of GWID for agent 17 during 2004 is not well captured by the tree model. This can be explained that the factors that lead to the spike of GWID are not considered by the tree model and future investigations need to be conducted to find out these factors and mitigate the discrepancy between the observation and the simulation.

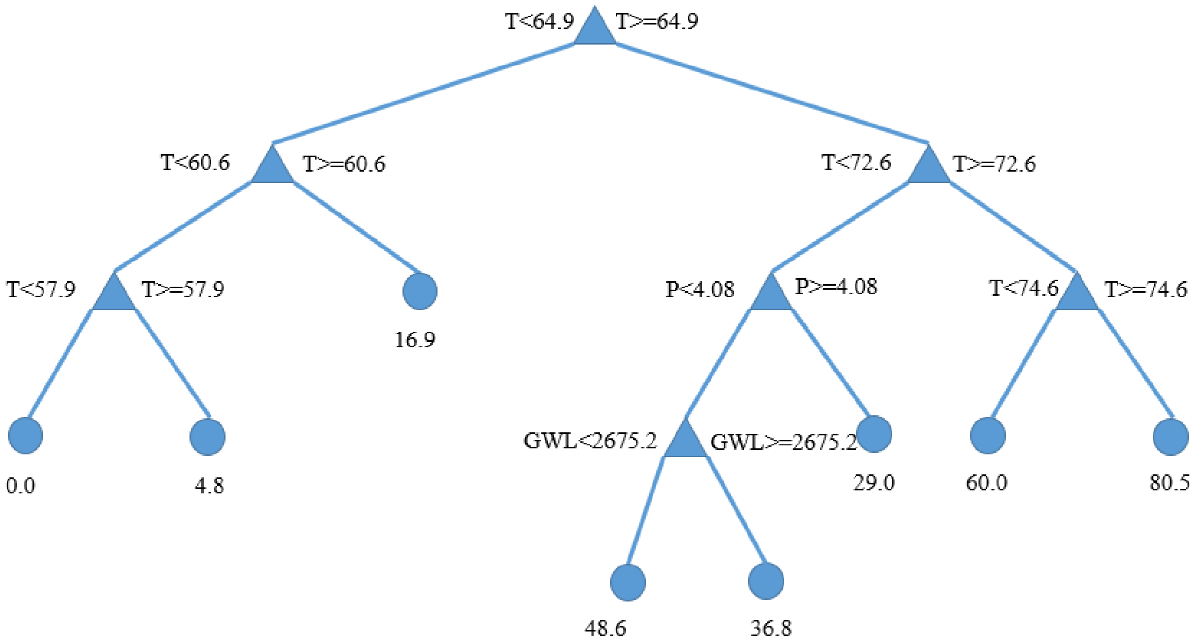


Figure 4-8 Tree No. 1 of the 500 trees for agent 24; T: monthly mean temperature [ºF]; GWL: groundwater level [ft]; P: monthly mean precipitation [inch]; solid black nodes are monthly groundwater irrigation depth [mm].

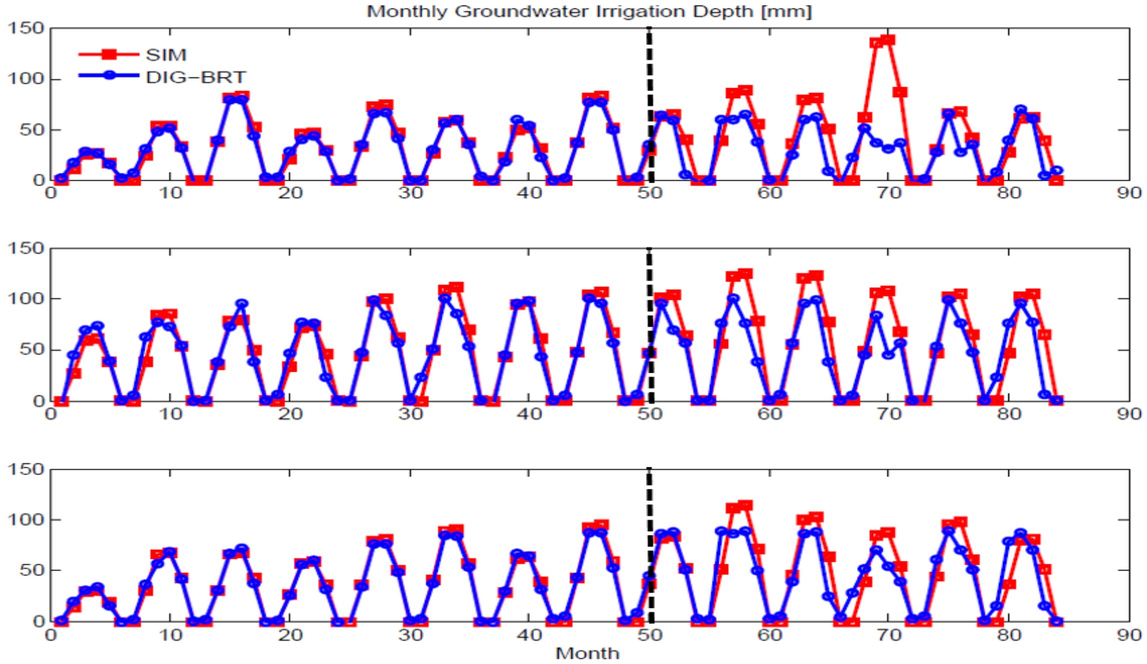


Figure 4-9 Comparisons of monthly groundwater irrigation depth between the observation (red) and the model simulation using boosted regression tree model (blue) for agents 17, 18 and 24 (from top to bottom) between year 1993~2006. The dashed line separates the training datasets from the validation ones.

All GWIDs for individual agents are then converted to groundwater monthly pumping rate, which is used as the driving force to the RRCA model as shown in Figure 4-10, and water table is then simulated through the RRCA model. The coupled ABM and RRCA models are used to investigate the impacts of agents' pumping behavior on the underlying groundwater system. We ran the coupled models from year 1993 to 2006, namely the directed information graph-boosted regression tree (DIG-BRT) scenario.

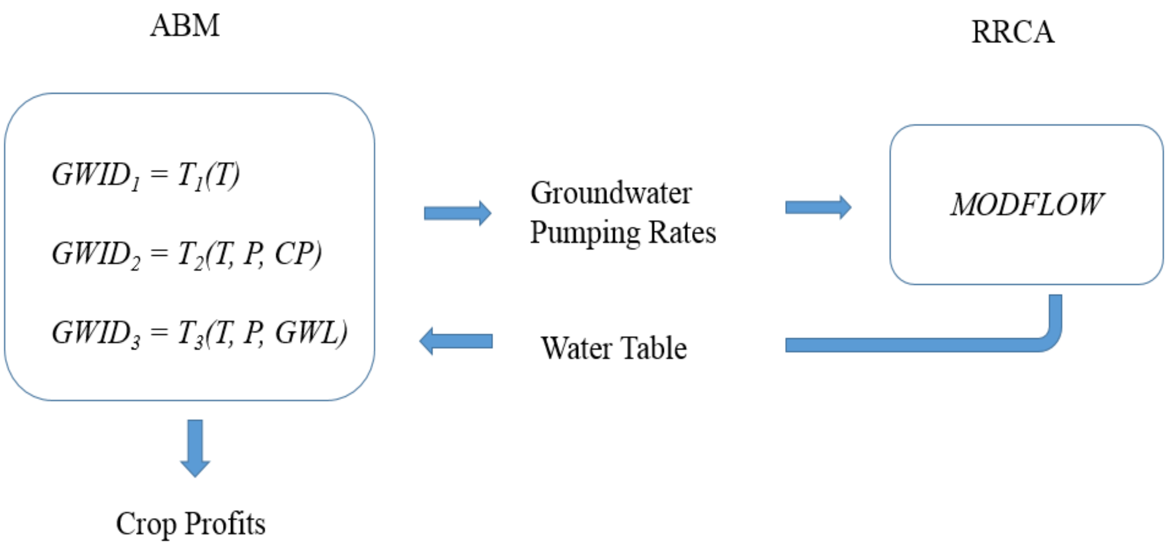


Figure 4-10 Coupling of agent-based model (ABM) with the groundwater model (RRCA); three representative graphs are simulated by the corresponding BRT models denoted by T_1, T_2 and T_3. The simulation period is from year 1993 to 2006.

The agents' behavioral rules derived using DIG and BRT attempt to mimic actual agents' behavioral rules. Figure 4-11 (a) and (b) show the comparisons of crop profits and water tables between the simulation scenario (using the input data from the RRCA model) and the DIG-BRT scenario for agents 17, 18, 24 whose behavioral rules are represented by one of the three ensemble tree models respectively shown in Figure 4-10. The results for different agents from the DIG-BRT scenario match the ones from the simulation scenario well. In addition, we also compare crop profits and water tables for the rest agents under these two scenarios, and similar results are found-for the rest agents, crop profits and water tables under the DIG-BRT scenario can well resemble the simulation results. We think the high goodness of fit is the result of the combined effects of DIG and BRT: the former identifies the important variables that causally influence agents' behavior and the latter derives the good models to quantitatively describe the causal relationships.

Figure 4-11 (a) and (b) show the impact of variations of agents' behavior on crop profits and water table using the optimization-based approach, which simulates agents' behavior with behavioral parameters (Please refer to Hu, et al. (2015) for more details). The shaded area indicates the confidence intervals of crop profits and water table for agent 17, 18 and 24 as the result of 1,000 model evaluations with different values of behavioral parameters and the dashed line shows their mean values. Although the mean values from the optimization-based approach well mimic the crop profits and water tables from the simulation scenario for agent 24, these mean values either underestimate or overestimate the results from the simulation for agent 17 and 18. In contrast with the inconsistent performance of the optimization-based approach, the crop profits and water table as the result of the behavioral rules derived using DIG and BRT can well mimic the results from the simulation. In this sense, we can conclude that the data-driven approach using DIG and BRT outperforms the optimization-based approach to capture the uncertainty of agents' behavior as the result of bounded rationality and simulate the actual agents' behavior.

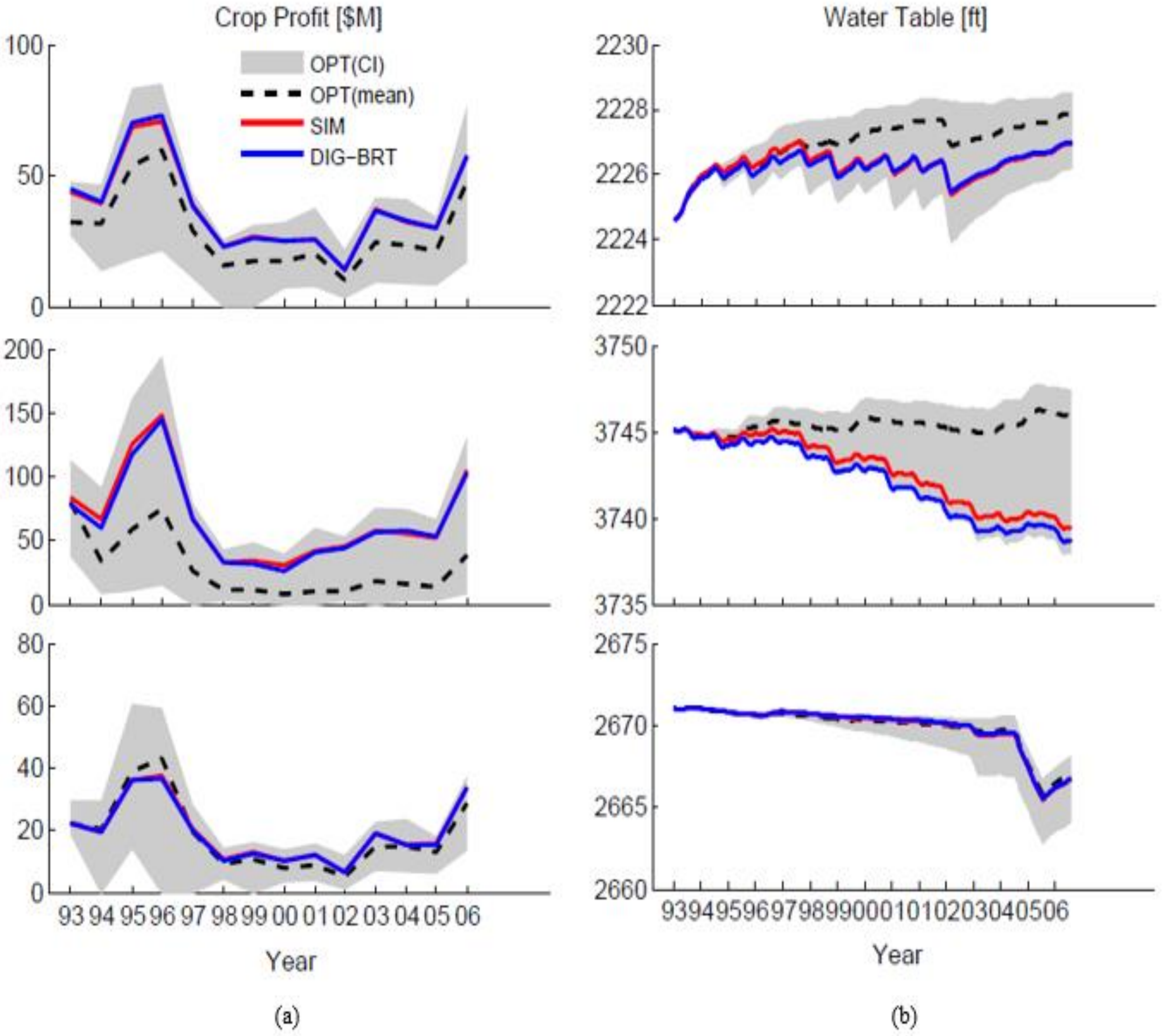


Figure 4-11 Comparisons of crop profits ($M; a) and water table (ft, b) between the simulation scenario (red), the DIG-BRT scenario (blue) and the optimization scenario (the shaded area is the confidence interval and the dashed line is the mean value) for agent 17, 18 and 24 (from top to bottom).

4.5 Summary and Conclusions

The most challenging aspect of agent-based modeling is to derive the agents' behavioral rules under the behavioral uncertainty, which arises from the fact that agents have bounded rationality in their decision making processes. In this chapter, we introduced a data-driven approach using the DIG as a vehicle to find the causal relationships between processes for the agent-based modeling in water resources management. Based on the measurement of directed information between variables that are likely to affect agents' behavior (i.e., groundwater irrigation depth in our case), we derive the corresponding DIG for each individual agent using the directed information graph algorithm. Expert domain knowledge and cross-validation are employed in the algorithm to rule out spurious causal relationships that could be caused by missing confounders, insufficient sample frequency, or noisy data. Through the combination of human and machine intelligence, we can derive the behavioral rules to describe agents' behavior as well as account for their bounded rationality. Based on the results, not only is it important to find out that the environmental factors like temperature and precipitation can affect agents' decision on groundwater irrigation depth, it is also interesting to show that the local irrigation practices can affect the key components of local climate, such as temperature and precipitation.

Based on the similarity of the DIG for each agent, three representative graphs are identified to represent all agents' behavior rules in the study area. We found that corn price, underlying groundwater level and monthly mean precipitation have causal influences on agents' decisions on groundwater irrigation depth to various extents; monthly mean temperature is the most common factor that affects all agents' irrigation behavior, especially in dry areas where temperature becomes the most dominant factor. Thus, our findings confirm that agents' irrigation behavior is consistent with the actual crop irrigation requirements, especially in dry areas.

An agent-based model is designed with behavioral rules characterized by three representative graphs, and coupled with the physically based groundwater model, the RRCA model. Through the coupled models, we investigate the impacts of agents' pumping behavior on the underlying groundwater system in the High Plains aquifer HO area. It is found that in comparison with the optimization-based approach, crop profits and water tables as the result of agents' pumping behavior derived using DIG and BRT can better mimic the actual ones from the simulation scenario. Thus, we can conclude that the data-driven approach using DIG and BRT outperforms the optimization based approach to capture the uncertainty of agents' behavior as the result of bounded rationality and mimic their actual behavior.

References

Aalders, I. (2008). Modeling land-use decision behavior with bayesian belief networks. *Ecology and Society, 13*(1), 16.

Alexandridis, K. T. (2006). *Exploring complex dynamics in multi agent-based intelligent systems: Theoretical and experimental approaches using the multi agent-based behavioral economic landscape (MABEL) model* ProQuest.

Amblard, P., & Michel, O. J. (2011). On directed information theory and granger causality graphs. *Journal of Computational Neuroscience, 30*(1), 7-16.

Anderson, J. R. (2007). How can the human mind occur in the physical universe? New York, NY: Oxford University Press

Arthur, W. B. (1999). Complexity and the economy. *Science (New York, N.Y.), 284*(5411), 107-109.

Betts, A. K. (2004). Understanding hydrometeorology using global models. *Bulletin of the American Meteorological Society, 85*(11), 1673-1688.

Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences of the United States of America, 99*, 7280-7287. doi:10.1073/pnas.082080899

Buntine, W. L. (1996). A guide to the literature on learning probabilistic networks from data. *Knowledge and Data Engineering, IEEE Transactions on, 8*(2), 195-210.

Carmona, G., Varela-Ortega, C., & Bromley, J. (2011). The use of participatory object-oriented bayesian networks and agro-economic models for groundwater management in Spain. *Water Resources Management, 25*(5), 1509-1524. doi:10.1007/s11269-010-9757-y

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd International Conference on Machine Learning,* 161-168.

Chase, T. N., Pielke, R. A., Kittel, T. G., Baron, J. S., & Stohlgren, T. J. (1999). Potential impacts on colorado rocky mountain weather due to land use changes on the adjacent great plains. *Journal of Geophysical Research, 104*(D14), 16.

Chickering, D. M., Geiger, D., & Heckerman, D. (1994). *Learning Bayesian networks is NP-hard* (Vol. 196). Technical Report MSR-TR-94-17, Microsoft Research.

Cover, T. M., & Thomas, J. A. (2012). *Elements of information theory* John Wiley & Sons.

Crooks, A. T., & Heppenstall, A. J. (2012). Introduction to agent-based modelling. *Agent-based models of geographical systems* (pp. 85-105) Springer.

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology, 77*(4), 802-813.

Elsawah, S., Guillaume, J. H., Filatova, T., Rook, J., & Jakeman, A. J. (2015). A methodology for eliciting, representing, and analysing stakeholder knowledge for decision making on complex socio-ecological systems: From cognitive maps to agent-based models.*Journal of Environmental Management, 151*, 500-516

Eltahir, E. A. (1998). A soil moisture-rainfall feedback mechanism 1. theory and observations. *Water Resources Research, 34*(4), 765-776.

Eltahir, E. A., & Bras, R. L. (1996). Precipitation recycling. *Reviews of Geophysics, 34*(3), 367-378.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society,* , 424-438.

Haddeland, I., Lettenmaier, D. P., & Skaugen, T. (2006). Effects of irrigation on the water and energy balances of the colorado and mekong river basins. *Journal of Hydrology, 324*(1), 210-223.

Harbaugh, A. W. (2005). *MODFLOW-2005, the US geological survey modular ground-water model: The ground-water flow process* US Department of the Interior, US Geological Survey Reston, VA, USA.

Hu, Y., Cai, X., & DuPont, B. (2015). Design of a web-based application of the coupled multi-agent system model and environmental model for watershed management analysis using hadoop. *Environmental Modelling & Software, 70*, 149-162.

Hu, Y., Garcia-Cabrejo, O., Cai, X., Valocchi, A. J., & DuPont, B. (2015). Global sensitivity analysis for large-scale socio-hydrological models using hadoop. *Environmental Modelling & Software, 73*, 231-243.

Janssen, M. A., & Ostrom, E. (2006). Empirically based, agent-based models. *Ecology and Society, 11*(2), 37.

Jordan, M. I. (1998). *Learning in graphical models:[proceedings of the NATO advanced study institute...: Ettore mairona center, erice, italy, september 27-october 7, 1996]* Springer.

Kanamitsu, M., & Mo, K. C. (2003). Dynamical effect of land surface processes on summer precipitation over the southwestern united states. *Journal of Climate, 16*(3), 496-509.

Kennedy, W. G. (2012). Modelling human behaviour in agent-based models. *Agent-based models of geographical systems* (pp. 167-179) Springer.

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques* MIT press.

Kustu, M. D., Fan, Y., & Robock, A. (2010). Large-scale water cycle perturbation due to irrigation pumping in the US high plains: A synthesis of observed streamflow changes. *Journal of Hydrology, 390*(3), 222-244.

Lloyd, S. P. (1982). Least squares quantization in PCM. *Information Theory, IEEE Transactions on, 28*(2), 129-137.

Loewenstein, G., & Lerner, J. S. (2003). The role of affect in decision making. *Handbook of Affective Science, 619*(642), 3.

Marko, H. (1973). The bidirectional communication theory--a generalization of information theory. *Communications, IEEE Transactions on*, *21*(12), 1345-1351.

McKusick, V. (2003). Final report for the special master with certificate of adoption of rrca groundwater model. *State of Kansas v.State of Nebraska and State of Colorado, in the Supreme Court of the United States, 3605*

Newell, A. (1972). A theoretical exploration of mechanisms for coding the stimulus. In A.W. Melton and E. Martin (Eds.), *Coding processes in human memory* (pp. 373–434). New York: Wiley.

North, M. J., & Macal, C. M. (2007). *Managing business complexity: Discovering strategic solutions with agent-based modeling and simulation* Oxford University Press.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference* Morgan Kaufmann.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Dubourg, V. (2011). Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research, 12,* 2825-2830.

Pielke, R. A. (2001). Influence of the spatial distribution of vegetation and soils on the prediction of cumulus convective rainfall. *Reviews of Geophysics, 39*(2), 151-177.

Quinn, C. J., Coleman, T. P., Kiyavash, N., & Hatsopoulos, N. G. (2011). Estimating the directed information to infer causal relationships in ensemble neural spike train recordings. Journal of Computational Neuroscience, 30(1), 17-44.

Quinn, C.J., Kiyavash, N., Coleman, T.P. (2015). Directed Information Graphs, in *Information Theory, IEEE Transactions on* , in press.

Reforgiato, D., Gutierrez, R., & Shasha, D. (2008). Graphclust: A method for clustering database of graphs. *Journal of Information & Knowledge Management, 7*(04), 231-241.

Robinson, D. T., Brown, D. G., Parker, D. C., Schreinemachers, P., Janssen, M. A., Huigen, M., . . . Irwin, E. (2007). Comparison of empirical methods for building agent-based models in land use science. *Journal of Land use Science, 2*(1), 31-55.

Simon, H. A. (1996). *The sciences of the artificial* (Vol. 136). MIT press.

Smajgl, A., Brown, D. G., Valbuena, D., & Huigen, M. G. (2011). Empirical characterisation of agent behaviours in socio-ecological systems. *Environmental Modelling & Software, 26*(7), 837-844.

Stanilov, K. (2012). Space in agent-based models. *Agent-based models of geographical systems* (pp. 253-269) Springer.

Tesfatsion, L. (2006). Agent-based computational economics: A constructive approach to economic theory. *Handbook of Computational Economics, 2*, 831-880.

Tze Ling Ng, Eheart, J. W., Cai, X., & Braden, J. B. (2011). An agent-based model of farmer decision-making and water quality impacts at the watershed scale under markets for carbon allowances and a second-generation biofuel crop. *Water Resources Research,47*, W09519. doi:10.1029/2011WR010399

Urban, C., & Schmidt, B. (2001). PECS-Agent-Based Modelling of Human Behaviour. In *Emotional and Intelligent II-The Tangled Knot of Social Cognition, AAAI Fall Symposium*.

Vörösmarty, C. J., & Sahagian, D. (2000). Anthropogenic disturbance of the terrestrial water cycle. *Bioscience, 50*(9), 753-765.

Wiener, N. (1956). The theory of prediction. *Modern Mathematics for Engineers, 1*, 125-139.

Yang, Y. E., Cai, X., & Stipanović, D. M. (2009). A decentralized optimization algorithm for multiagent system–based watershed management. *Water Resources Research, 45*(8).

5 CHAPTER V – FINAL REMARKS

This dissertation focuses on expanding the conventional environmental models with human factors for watershed management analysis in the context of coupled human and natural systems. Using an agent-based modeling framework, optimization-based and data-driven approaches are applied to modeling farmers' pumping decision-making processes in the High Plains aquifer within the hydrological observatory area. Please refer to the following subsections for a summary of major findings and discussions, as well as limitations of the study and possible future work.

5.1 Summary

Agent-based model (ABM) or multi-agent system (MAS) framework is used to model farmers' pumping decision-making processes in the High Plains aquifer within the hydrological observatory area. For my research, I developed different approaches (i.e., optimization-based and data-driven approaches) to design ABMs and coupled them with a physically-based groundwater model to investigate the interactions between farmers and the underlying groundwater system.

For the first part of my study, Chapter II, all agents are assumed to be rational. An optimization-based approach, which incorporates self-learning and utility maximization, is developed to simulate agents' decisions on crop types, optimal irrigated/dryland area, and irrigation depth at the annual time scale. The agent-based model (ABM) is then coupled with a physically-based groundwater model, RRCA model. Unfortunately, high computational intensity arises from the coupled models, which limits the applications of the models. As a result, multithreaded programming is used to ease the computation intensity of the coupled models by running agents in parallel. The result shows that the total running time of the coupled models is reduced by 80%, from one hour down to twelve minutes on an eight-core machine node.

A web-based application is built to provide network access to the coupled ABM and groundwater model. In order to ensure the web application of the coupled models with system and user scalability, a framework which combines multithreaded programming with Hadoop-based cloud computing is developed. The multithreaded programming is applied to improving the computational efficiency of the single instance of the model (i.e. system scalability); Hadoop-based cloud computing provides on-demand computational power to execute multiple instances of the model simultaneously (i.e. user scalability). As a result, this part of my work presents an initial effort of modeling the coupled human behavioral model and environmental model as a web application, which can facilitate an online dissemination of the models, and support participatory modeling exercises.

The optimization-based approach simulates agents’ behavior with behavioral parameters. These behavioral parameters affect agents' predictions of the future crop prices and precipitation through the learning process, which are used later by agents to determine the optimal pumping rates while maximizing their utilities. For the second part of the study, Chapter III, we investigated how these behavioral parameters affect the outcomes of the coupled models in terms of crop profits and water tables using global sensitivity analysis (GSA), which provides us with better understanding of the interactions between farmers’ pumping behavior and the underlying groundwater systems.

A large number of model evaluations is required for GSA. Thus, two approaches are developed to address the computational issues arising from the execution of the coupled models using the MapReduce framework. Approach I exploits both the independence of model evaluations and the lack of explicit interactions between agents, and thus different agents in various scenarios can run their tasks simultaneously with different machine nodes. Different from Approach I, Approach II exploits only the independence of model evaluations, and thus different scenarios (rather than different agents) are executed with different machine nodes. Through the analysis, it is found that the second approach is more suitable for our case with limited computational resources. As a result, with Approach II, a substantial reduction of the computation time is achieved, from 42 days required to run 1,000 scenarios sequentially on a desktop machine to two hours by running them on the Illinois Cloud Computing Testbed.

With the well-represented sample sets of the behavioral parameters and the mechanism to efficiently run the coupled models 1,000 times, a large amount of crop profits and water table data are generated. The PCE is applied to generating the surrogate model for the complex coupled models using the data sets. The variance-based sensitivity indices are then calculated using the PCE coefficients. As a result, GSA using PCE-based variance decomposition approach identifies the influential parameters (i.e., κ_{pr} and λ) and quantify the spatio-temporal interactions between agents and the groundwater system through these parameters.

Overall, for Chapter III, we developed a methodology framework for the application of GSA to the coupled models. This framework attempts to find a balance between the heavy computational burden associated with the model execution and the number of model evaluations required for GSA analysis. The balance is achieved through the combination of Hadoop-based cloud computing and Polynomial Chaos Expansion (PCE); the former can efficiently execute a large number of complex models in parallel and the latter allows efficient estimation of sensitivity indices from PCE coefficients.

However, the optimization-based approach is developed based on the assumption that agents are rational. But, in fact, agents have bounded rationality in their decision-marking processes. For the third part of this dissertation, Chapter IV, a data-driven approach is thus introduced to derive agents’ behavioral rules considering the influence of bounded rationality. The DIG approach is used as a vehicle to find the causal relationships between variables that that are likely to affect agents’ behavior (i.e., groundwater irrigation depth in our case) as well as reflect agents’ bounded rationality. Expert domain knowledge and cross-validation are employed in our approach to rule out spurious causal relationships that could be caused by missing confounders, insufficient sample frequency, or noisy data.

Each agent is associated with a DIG that describes their pumping behavior. We found that corn price, underlying groundwater level and monthly mean precipitation have causal influences on agents’ decisions on groundwater irrigation depth to various extents; monthly mean temperature is the most common factor that affects all agents’ irrigation behavior in the study area, especially in dry districts where temperature becomes the most dominant factor. Based on the similarity of their graphs, three representative graphs are identified to represent all agents’ behavior rules in the study area.

An agent-based model is designed with behavioral rules characterized by three representative graphs using boosted regression tree (BRT) models, and coupled with the physically-based groundwater model, the RRCA model. Through the coupled models, the investigation is conducted to understand the impacts of agents' pumping behavior on the underlying groundwater system in terms of crop profits and water tables. In comparison with the results from the optimization-based approach, it is found that crop profits and water tables as the result of agents' pumping behavior derived using the data-driven approach can better mimic the actual ones from the simulation scenario (i.e., use the input data from RRCA model). Thus, it is concluded that the data-driven approach based on DIG and BRT outperforms the optimization-based approach to capture the uncertainty of agents' behavior as result of bounded rationality.

Overall, I have demonstrated two different approaches to design agent-based model to model famers' pumping behavior in my dissertation, optimization-based and data-driven approaches, and shown that both approaches have their own limitations: the optimization-based approach usually assumes the rationality of agents. It also requires modelers to have a good understanding of the mechanism that drives agents to behave, and model these behaviors with behavioral parameters. Although these limitations can be dealt with by the data-driven approach, its applications are often constrained by the data availability. In the following section, I will delve into the assumptions made for the ABM development as the result of data shortage in my work, and suggest how to address these assumptions when data become available in the future

5.2 Limitations and Future Work

The major challenge of agent-based modeling is to derive agents' behavioral rules due to limited empirical data. Assumptions made in this dissertation need to be verified by data when they become available. In the following, I would like to review the major assumptions made in the dissertation, describe limitations as the result of the assumptions and discuss strategies for future work to tackle with these limitations.

Agents can interact with each other in various ways. In my current work, only the indirect interaction between agents through the RRCA model is considered, that is, the impact of one agent's pumping behavior affects other agents through the shared underlying groundwater resource. However, in the real world, there can exist direct interactions between agents. For example, the adaptation of new irrigation technologies by one agent can promote the adaptation of these technologies by other agents. The current agent-based model is not able to capture such direct interactions, which is mainly due to the lack of empirical data to describe direct interactions between agents.

In Chapter II, agents' pumping behavior is derived from the optimization-based approach which assumes that all agents optimize their utility through learning the future crop prices and precipitation. Agents' learning process is simulated using Bayesian statistics. However, little evidence can be used to justify the way to simulate agents' beliefs in future crop prices and precipitation as a Bayesian learning process, or even more fundamental, if and how real famers incorporate their beliefs into their decision-making processes in the real world. Again, these assumptions have been made due to the lack of empirical data that can be used to update the current learning and optimization strategies with more realistic ones.

Moreover, agents' beliefs in prices or precipitation can be correlated across county lines. Presumably, agents' rely on similar information sources (e.g., weather reports and radio) would shift their beliefs in a correlated fashion. However, no data is available to measure the correlation between agents' preferences across county lines. Thus, in Chapter III, all five behavioral parameters defined for each agent to describe their preferences between prior knowledge and historical experience of crop prices and precipitation are assumed to follow uniform distribution. For the future work, it will be useful to collect data to verify the distributions of their behavioral parameters and how agents' beliefs are correlated through these parameters.

Unlike measuring physical quantities, ways to measure human behavior directly are limited, especially for emotional and social aspects of human behavior. Conventionally, researchers use social surveys or interviews to gather human behavioral data indirectly. However, the data quality is subject to uncontrolled human factors, such as the design of questionnaire, interviewers' skills and their interactions with the respondents. Fortunately, thanks to the new technologies, such as portable sensors, wearable devices and social media, we are provided with various means to directly measure human activities and their interactions with environmental systems. Thus, future work will include the collection of data related to farmers' decision-process on pumping. With these data, the aforementioned assumptions will be carefully checked and updated. In addition to data collections, other work can be done in parallel.

In Chapter II, a relational database is used for data management of the web application of the coupled models, and running many instances simultaneously can exceed the constraint on the number of database connections set by the provider, which affects the user scalability. In the future work, a distributed NoSQL database could be implemented to replace the current relational database. As a result, the simultaneous execution of the coupled models should scale linearly with the number of nodes available on the cluster. In addition, when agents' direct interactions are included in the set of behavioral rules, the computational design for the parallel implementation of agent-based model should be updated accordingly.

In Chapter III, the major focus is the development of a computational framework to identify the most influential behavioral parameters. For the future work, it is worthwhile to understand why the variations of these behavioral parameters have significant impacts on the coupled models, and explore the implications of the results of temporal and spatial sensitivity analysis for the design of sustainable groundwater use policy in the study area. It will be also interesting to apply the auto-calibration techniques (e.g. simulated annealing) to calibrating the coupled models against the observation data to find the optimal values of the influential behavioral parameters.

In Chapter IV, the results show that the behavioral rules derived from data-driven approach outperforms the optimization-based approach to simulate agents' pumping behavior in the real world. However, there are still some parts of agents' behavior (i.e. monthly groundwater irrigation depth) are not well captured by the rules from the data-driven approach. For example, agents tend to use more water in the real world than that predicted by our approach, and it will be worthwhile to make efforts to address the prediction discrepancy.

Finally, this dissertation mainly investigates the impacts of agents' pumping behavior on the underlying groundwater systems in terms of water tables. For the future work, it will be meaningful to explore how agents' behavior affect surface water and ecosystems via base flow, i.e., water exchange between groundwater and surface water.

Appendix A

A.1 Optimization (utility maximization)

This section will give an overview of the optimization approach used to describe agents' decision making processes. We start with the crop yield function (Palazzo, 2009).

A.1.1 Crop yield function

$$Y = Y_d + (Y_m - Y_d)[1-(1-I_r)^{1/\beta}]. \tag{A.1}$$

Where Y : crop yield [bu/acre]; Y_m : maximum crop yield without water stress [bu/acre]; Y_d : rainfed yield [bu/acre]; β : irrigation efficiency ($\beta \in [0,1]$), calculated by $(\mathrm{ET}_m\text{-}\mathrm{ET}_d)/\mathrm{I}_m$; ET_m : seasonal evapotranspiration under no water stress [inches]; ET_d : seasonal evapotranspiration for dryland [inches], which is defined as the sum of effective rainfall, P_e [inches/month] and monthly allowable soil water depletion during the crop growing season, W_b [inches/month] : $\mathrm{ET}_d = P_e + W_b$, in which we assume that no soil water depletion occurs during the crop growing season, equivalent to $W_b = 0$. Thus, P_e is equal to ET_d ; I_m is the maximum water demand [inches]; I_r : the ratio of actual irrigation depth to the maximum irrigation depth denoted by $\mathrm{I/I}_m$.

A.1.2 Estimation of effective rainfall

Effective rainfall means useful or utilizable rainfall, which is defined as the portion of rainfall used for crop growth. Effective rainfall does not include the rainfall allocation producing runoff or drainage below the root zone. The monthly potential effective rainfall is estimated using the USDA-SCS method (USDA, 1967):

$$P_e^* = max[0, (1.253 \cdot \tilde{p}_m^{0.824} - 2.935) \cdot 10.0^{(0.001ET_c)}]$$
$$P_e = \begin{cases} min(P_e^*, \tilde{P}_m, ET_c) & \tilde{p}_m > 12mm \\ \tilde{p}_m & Otherwise. \end{cases} \tag{A.2}$$

Where $\tilde{P}_m$ is mean monthly rainfall [mm/month]; ET_c is the potential crop evapotranspiration in the absence of environmental or water stress [mm/month]. In the crop coefficient approach the crop evapotranspiration, ET_c is calculated as the product of the reference crop evapotranspiration, ET_0 and the crop coefficient, K_c: $ET_c = K_c \cdot ET_0$. Here, we calculate daily ETo using the so-called 1985-Hargreaves equation (HG) (Hargreaves and Samani, 1985): $ET_o = 0.0023 \cdot (T_{mean} + 17.8)(T_{max} - T_{min})^{0.5} Ra$; $\mathrm{T}_{\max}$: the maximum daily air temperature [°C], $\mathrm{T}_{\min}$: the minimum daily air temperature [°C]; R_a: the extraterrestrial solar radiation [$MJ \cdot m^{-2} \cdot d^{-1}$].

Further, we assume the maximum yield is reached when ET_c is equal to ET_m. Then, we obtain the crop yield function for different crops

$$Y_j = Y_{d,j} + (Y_{m,j} - Y_{d,j})[1 - (1 - \beta \cdot I_j / (K_{c,j} \cdot ET_o - P_{e,j}))^{1/\beta}]. \quad (A.3)$$

and the crop profit function π is given as:

$$\pi = \sum_{j=1}^{4} \pi_j = \sum_{j=1}^{4} (A_{i,j}[(P_{i,j} - t_{i,j}) \cdot Y_j - cw_{i,j} \cdot I_j - F_{i,j}] + A_{d,j}[(P_{i,j} - t_{i,j}) \cdot Y_{d,j} - F_{d,j}]). \quad (A.4)$$

Where i and d represent irrigated and dryland (rainfed) area respectively. $A_{i,j}$ and $A_{d,j}$ are the irrigated area and dryland area [acre]; $P_{i,j}$ and $P_{d,j}$ are the crop prices for crop j in the irrigated area and dryland area [$/bu]; $t_{i,j}$ is the corresponding transportation cost [$/bu]; $F_{i,j}$ and $F_{d,j}$ are the crop-specific fixed cost of production [$/acre]. Different from planting in rainfed area, $cw_{i,j}$ is the energy cost associated with pumping [$/arce-inch]. The energy cost function for pivot irrigation is given as follows (Palazzo, 2009):

$$\begin{aligned} cw_{i,j}(\hat{h}) = {} & \frac{C \cdot T \cdot E}{A_{i,j} \cdot D_{i,j}} + \frac{Lh \cdot Cl}{(3 - 1.5 \cdot Pivot) \cdot A_j} + \\ & 9 \times 10^{-6} \cdot (\hat{h} + 2.31 \cdot P) - 2.4 \times 10^{-3} \cdot (\hat{h} + 2.31 \cdot P)^2 + 2.9137. \end{aligned} \quad (A.5)$$

Where $\mathrm{cw}_{i,j}$ is the energy cost associated with pumping for crop j planted in the irrigated area denoted by i [$/arce-inch]; C is the cost of diesel fuel [$/gallon]; D_j is the water allocation depth [inches]; T is the total operation time of pumps [h], given by $7.48\times 43560/(12\times 60v)\cdot A_{i,j}\cdot D_{i,j}$, where v is the pumping rate [gpm]; E is the hourly energy consumption [gallon/h]; Lh is the number of labor hours required per irrigation [h]; Cl is the unit labor cost [$/h]; Pivot denotes the irrigation depth using the pivot irrigation system [inch]; $\hat{h}$ is the distance that water must be lifted to the elevation of pump discharge[feet]; we assume $\hat{h}$ equals the distance from the land surface to the groundwater level; $\hat{h}$ is computed via the centroid weighted average using the inverse of the distance of each pumping well within the agent to the centroid of the agent as the weight; P is the pump discharge pressure [psi]. Notice that as the water table lowers, agents will have to bear higher energy costs.

A.1.3 Two-stage optimization

In order to mimic an agent's actual pumping decision process, we decompose the utility maximization problem defined by Equation 1 in Chapter II into two sub-problems using a two-stage optimization strategy. First, before the crop planting and growing season agents make predictions of the future crop prices and precipitation for the entire planting and growing season. The aforementioned Bayesian learning process is used to simulate agents' learning from their prior knowledge and experiences, and update their predictions of the posterior distribution of crop prices and precipitation, and expected water demand [inches], $I_{e,j}$ for crop j accordingly. Given the posterior distribution of the crop prices and precipitation, a sampling-based stochastic optimization method is developed to simulate agents' decisions on the choices of crop type, j and the corresponding planted irrigated and rainfed crop area [acre], $A_{i,j}^{planted}$ and $A_{d,j}^{planted}$. Thus, the utility maximization approach for the first-stage optimization strategy is formulated as follows:

$$\begin{aligned} &\text{maximize} && U(\pi) \\ &\textit{subject to} && \sum_{j=1}^{4}(A_{i,j}^{planted}+A_{d,j}^{planted})\leq \bar{A}, \\ & && \sum_{j=1}^{4}A_{i,j}^{planted}I_{e,j}\leq TWA, \\ & && A_{i,j}^{planted}\geq 0, A_{d,j}^{planted}\geq 0. \end{aligned} \tag{A.6}$$

Where $\bar{A}$ is the total arable land [acre]; TWA denotes the estimated total water availability for each individual agent [acre-inch].

For the second stage, agents determine the optimal water use I_j and irrigated and dryland area $\mathrm{A}_{i,j}$ and $\mathrm{A}_{d,j}$ given the observations of crop prices, precipitation at the current stage and the planted crop areas from the first stage. A deterministic optimization problem is formulated based on the crop profit function:

$$\begin{aligned} &\text{maximize} && \pi \\ &\textit{subject to} && \sum_{j=1}^{4} A_{i,j} I_j \le TWA, \\ & && 0 \le A_{i,j} \le A_{i,j}^{planted}, 0 \le A_{\mathrm{d},j} \le A_{i,j}^{planted} + A_{\mathrm{d},j}^{planted}, \\ & && 0 \le I_j \le I_{\mathrm{m},j}. \end{aligned} \tag{A.7}$$

Where $\mathrm{I}_{m,j}$ is the maximum crop water demand [inches].

Appendix B

B.1 Bayesian Learning

A Bayesian learning framework is used to simulate agents' ability to predict the future crop prices and precipitation during the crop growing season (Hu et al., 2015). The framework uses Bayesian statistics to incorporate the observations of crop prices and precipitation before planting the crops (i.e., simulated as likelihood functions) into their past experiences of them (i.e., prior knowledge) to update their predictions of crop prices and precipitation (i.e., posterior knowledge). For crop prices and precipitation, we assume that their likelihood functions follow the normal distribution:

$$p(D_{obs} \mid \mu,\sigma^2) = \frac{1}{(2\pi)^{n/2}}(\sigma^2)^{-n/2}\exp(-\frac{1}{2\sigma^2}[n\sum_{i=1}^{n}(x_i - \overline{x}) + n(\overline{x} - \mu)^2]) \qquad \text{(B.1)}$$

where $D_{obs} = (x_1,\cdots,x_i,\cdots,x_n)$ are the observations, the sequence of which is independent and identically distributed (IID) and $\overline{x}$ is the mean of the sequence. μ and σ^2 are the mean and variance of the likelihood function.

A suitable conjugate prior, the normal-inverse-chi-squared ($NI\chi^2$) prior as the product of normal distribution (N) and inverse-chi-squared distribution (χ^{-2}) is used (Murphy, 2007):

$$p(\mu,\sigma^2) = NI\chi^2(\mu_0,\kappa_0,\nu_0,\sigma_0^2) = N(\mu \mid \mu_0,\sigma^2/\kappa_0)\chi^{-2}(\sigma^2 \mid \nu_0,\sigma_0^2) \qquad \text{(B.2)}$$

where μ_0 is the prior mean and κ_0 is how strongly we believe the prior mean; σ_0^2 is the prior variance and ν_0 is how strongly we believe this. The hyperparameters μ_0 and σ^2/κ_0 can be interpreted as the location and scale of μ, and the hyperparameters ν_0 and σ_0^2 as the degrees of freedom and the scale of σ^2. Then, we obtain the posterior distributions of prices and precipitation via Bayes theorem (Lee, 2004, P67):

$$
\begin{aligned}
& p(\mu,\sigma^2 \mid D_{obs}) = NI\chi^2(\mu_n,\kappa_n,\nu_n,\sigma_n^2) \propto p(\mu,\sigma^2)\,p(D_{obs} \mid \mu,\sigma^2) \\
& \qquad \mu_n = \frac{\kappa_0\mu_0 + n\bar{x}}{\kappa_n} \\
& \qquad \kappa_n = \kappa_0 + n \\
& \qquad \nu_n = \nu_0 + n \\
& \qquad \sigma_n^2 = \frac{1}{\nu_n}\left(\nu_0\sigma_0^2 + \sum (\mathrm{x}_i - \bar{x})^2 + \frac{n\kappa_n}{n+\kappa_n}(\mu_0 - \bar{x})^2\right)
\end{aligned}
\tag{B.3}
$$

where μ_n is the posterior mean and κ_n represents the level of confidence to the posterior mean; σ_n^2 is the posterior variance and ν_n reflects the level of confidence to the posterior variance. As a result, agents update their annual predictions of the expected crop prices and precipitation given their new observations, which will further impact agents' decisions on groundwater pumping for irrigation (Hu et al., 2015).

B.2 Polynomial Chaos Expansion

Let's suppose that we have a mathematical model that can be expressed as $\mathrm{Y} = \mathbf{g}(\xi),$ where Y and ξ are the output and input random variables respectively, and ξ follows a uniform distribution. In this case, the PCE of the output variable Y is given in terms of the Legendre Polynomials Ψ of a degree that is defined by the complexity of the relationship between the input and output variables. An example of a PCE of Y using unidimensional Legendre polynomials Ψ up to degree four ($M = 1, p = 4$) is given by:

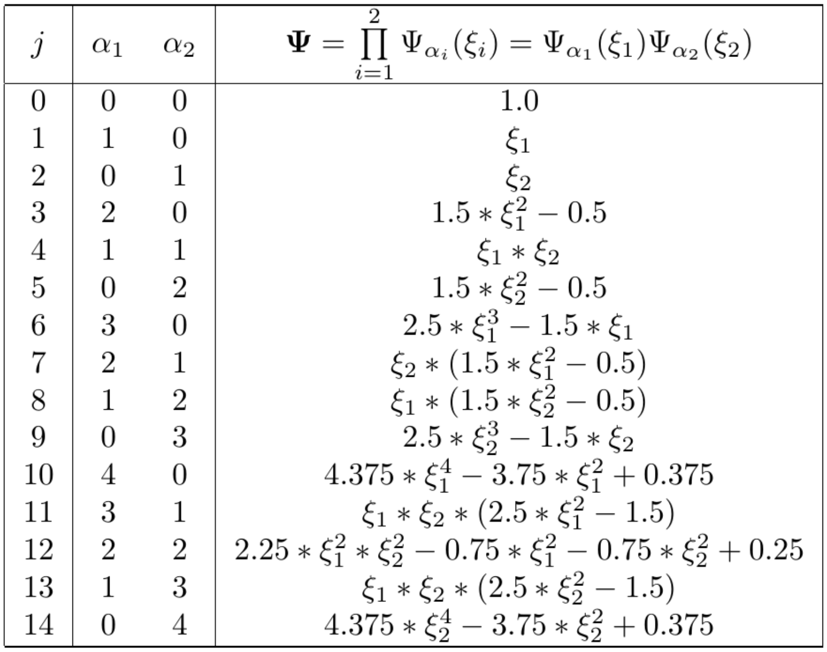
$$
\begin{aligned}
\mathrm{Y} = {} & \beta_0 \underset{\Psi_0}{1.0} + \beta_1 \underset{\Psi_1}{\xi} + \beta_2 \underbrace{\left(1.5\xi^2 - 0.5\right)}_{\Psi_2} + \beta_3 \underbrace{\left(0.5\xi\left(5.0\xi^2 - 3.0\right)\right)}_{\Psi_3} \\
& + \beta_4 \underbrace{(4.375\xi^4 - 3.75\xi^2 + 0.375)}_{\Psi_4} = \sum_{\mathbf{i}=0}^{4} \beta_{\mathbf{i}} \Psi_{\mathbf{i}}(\xi)
\end{aligned}
\tag{B.4}
$$

where $\beta_0, \beta_1, \beta_2, \beta_3$ are the coefficients that must be estimated from a set of N_R model evaluations $\{Y^{(ir)} = g(\xi^{(ir)})\}, ir = 1, \ldots, N_R$ using either linear regression (Sudret, 2008; Eldred and Burkardt,

2009) or integration methods taking advantage of the orthogonality of the polynomials (Ghanem and Spanos, 1991; Xiu and Karniadakis, 2002 and Xiu, 2010).

If the model $\mathrm{Y} = g(\xi_1, \xi_2)$ depends on two input variables $(M = 2)$ then the PCE of Y requires the use of multivariate orthogonal polynomials that can be defined via a tensor product of polynomials of a single variable. This product can be easily calculated using the integer set $\boldsymbol{\alpha}$ defined in Equation 7 in Chapter III that defines the specific terms that are multiplied. A simple example of the terms of the PCE of Y of order two $(M = 2)$ and using multivariate orthogonal polynomials $\boldsymbol{\Psi}$ of degree up to four $(p = 4)$ is given in the Table B-1 where a total of $D+1=(2+4)!/(2!4!)=15$ are required.

Table B-1: Number, integer set $\boldsymbol{\alpha}$ and multivariate orthogonal polynomials defined from the tensor product of the unidimensional Legendre Polynomials.



j	α_1	α_2	$\boldsymbol{\Psi} = \prod_{i=1}^{2} \Psi_{\alpha_i}(\xi_i) = \Psi_{\alpha_1}(\xi_1)\Psi_{\alpha_2}(\xi_2)$
0	0	0	1.0
1	1	0	ξ_1
2	0	1	ξ_2
3	2	0	$1.5 * \xi_1^2 - 0.5$
4	1	1	$\xi_1 * \xi_2$
5	0	2	$1.5 * \xi_2^2 - 0.5$
6	3	0	$2.5 * \xi_1^3 - 1.5 * \xi_1$
7	2	1	$\xi_2 * (1.5 * \xi_1^2 - 0.5)$
8	1	2	$\xi_1 * (1.5 * \xi_2^2 - 0.5)$
9	0	3	$2.5 * \xi_2^3 - 1.5 * \xi_2$
10	4	0	$4.375 * \xi_1^4 - 3.75 * \xi_1^2 + 0.375$
11	3	1	$\xi_1 * \xi_2 * (2.5 * \xi_1^2 - 1.5)$
12	2	2	$2.25 * \xi_1^2 * \xi_2^2 - 0.75 * \xi_1^2 - 0.75 * \xi_2^2 + 0.25$
13	1	3	$\xi_1 * \xi_2 * (2.5 * \xi_2^2 - 1.5)$
14	0	4	$4.375 * \xi_2^4 - 3.75 * \xi_2^2 + 0.375$

From the values in Table B-1, the PCE of Y using multivariate orthogonal polynomials is given by:

$$
\begin{aligned}
\mathrm{Y} &= \sum_{\mathbf{j}=0}^{\mathbf{D}} \beta_{\mathbf{j}} \boldsymbol{\Psi}_{\boldsymbol{\alpha}\mathbf{j}} \\
&= \beta_0 + \beta_1 \xi_1 + \beta_2 \xi_2 + \beta_3 (1.5\xi_1^2 - 0.5) + \beta_4 \xi_1 \xi_2 \\
&\quad + \beta_5 (1.5\xi_2^2 - 0.5) + \beta_6 (2.5\xi_1^3 - 1.5\xi_1) + \beta_7 \xi_2 (1.5\xi_1^2 - 0.5) \\
&\quad + \beta_8 (\xi_1 (1.5\xi_2^2 - 0.5)) + \beta_9 (2.5 * \xi_2^3 - 1.5 * \xi_2) \\
&\quad + \beta_{10} (4.375\xi_1^4 - 3.75\xi_1^2 + 0.375) + \beta_{11} (\xi_1 \xi_2 (2.5\xi_1^2 - 1.5)) \\
&\quad + \beta_{12} (2.25\xi_1^2 \xi_2^2 - 0.75\xi_1^2 - 0.75\xi_2^2 + 0.25) \\
&\quad + \beta_{13} (\xi_1 \xi_2 (2.5\xi_2^2 - 1.5)) + \beta_{14} (4.375\xi_2^4 - 3.75\xi_2^2 + 0.375)
\end{aligned}
\tag{B.5}
$$

where the coefficients $\{\beta_{\mathbf{j}}; \mathbf{j} = 0, \ldots, 14\}$ are estimated in the same way as the unidimensional case if a set of model evaluations is available.

Appendix C

C.1 Statistical Quantities

We now briefly introduce statistical quantities which are used in the algorithms (see Ch. 2 of Cover and Thomas, 2012). The *Kullback-Leibler* (KL) divergence,

$$\mathbf{D}(\mathbf{P} \,\|\, \mathbf{Q}) := \mathbf{E}_{\mathbf{P}}\left[\log\frac{\mathbf{P}(\mathbf{Z})}{\mathbf{Q}(\mathbf{Z})}\right] \tag{C.1}$$

measures how close the distribution **Q** is to **P**, where both are over the same random variable **Z**. The *mutual information* defined as

$$\mathbf{I}(\mathbf{X};\mathbf{Y}) := \mathbf{D}(\mathbf{P}_{\mathrm{X,Y}} \,\|\, \mathbf{P}_{\mathbf{X}}\mathbf{P}_{\mathbf{Y}}) \tag{C.2}$$

measures how correlated two random variables **X** and **Y** are by comparing their joint distribution to the product of the marginals. Equation A.2 is well defined for any distribution **P**. If **X** and **Y** are independent, $\mathbf{P}_{\mathbf{X,Y}}(\mathbf{x},\ \mathbf{y}) = \mathbf{P}_{\mathbf{X}}(\mathbf{x})\mathbf{P}_{\mathbf{Y|X}}(\mathbf{y}\,|\,\mathbf{x}) = \mathbf{P}_{\mathbf{X}}(\mathbf{x})\mathbf{P}_{\mathbf{Y}}(\mathbf{y})$ so $\mathrm{I}(\mathrm{X};\mathrm{Y}) = 0.$ Note that mutual information can be used to quantify the statistical correlation for complex, non-linear models.