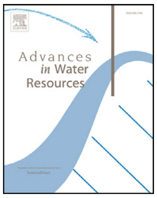
Advances in Water Resources 109 (2017) 29–40

Contents lists available at ScienceDirect

Advances in Water Resources

journal homepage: www.elsevier.com/locate/advwatres



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

Yao Hu[a,b,*], Christopher J. Quinn[c], Ximing Cai[a], Noah W. Garfinkle[a]

[a] Ven Te Chow Hydrosystems Laboratory, Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA
[b] Civil & Environmental Engineering Department, University of Michigan, Ann Arbor, MI, USA
[c] School of Industrial Engineering, Purdue University, West Lafayette, IN, USA

ARTICLE INFO

Article history:
Received 11 March 2017
Revised 15 August 2017
Accepted 16 August 2017
Available online 1 September 2017

Keywords:
Agent-based modeling
Bounded rationality
Behavioral uncertainty
Probabilistic graphical model
Directed information graph
Boosted regression trees

ABSTRACT

For agent-based modeling, the major challenges in deriving agents' behavioral rules arise from agents' bounded rationality and data scarcity. This study proposes a "gray box" approach to address the challenge by incorporating expert domain knowledge (i.e., human intelligence) with machine learning techniques (i.e., machine intelligence). Specifically, we propose using directed information graph (DIG), boosted regression trees (BRT), and domain knowledge to infer causal factors and identify behavioral rules from data. A case study is conducted to investigate farmers' pumping behavior in the Midwest, U.S.A. Results show that four factors identified by the DIG algorithm- corn price, underlying groundwater level, monthly mean temperature and precipitation- have main causal influences on agents' decisions on monthly groundwater irrigation depth. The agent-based model is then developed based on the behavioral rules represented by three DIGs and modeled by BRTs, and coupled with a physically-based groundwater model to investigate the impacts of agents' pumping behavior on the underlying groundwater system in the context of coupled human and environmental systems.

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 stakeholders' behavior, in particular their interactions with their biophysical systems, has never 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, gaining its popularity in both social science and economics (Arthur, 1999; Bonabeau, 2002; Tesfatsion, 2006).

The design of an agent-based model follows a bottom-up, distributed approach. It starts 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., 2015a). Employing ABMs 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, ABMs outweigh conventional simulation models, built based on the top-down centralized approach, in studying the system dynamics. ABMs are more likely to capture emergent phenomena arising from the interactions between human and environmental systems.

Modeling human behavior is complex. Human behavior is not random but based on our diverse knowledge and abilities, and modeling such behavior would not be particularly challenging if it were always rational (Kennedy, 2012). The rationality of human behavior is affected by emotional, intuitive, or unconscious decision-making processes. These processes can distort agents' perceptions of the environment and the likelihood of future evaluations (Loewenstein and Lerner, 2003). Furthermore, limited information, varying cognitive abilities and insufficient time all contribute to limit the rationality of human decision making (Simon, 1996). Regardless of its origin, agents' bounded rationality makes it difficult, if not impossible, to derive "perfect" rules for an ABM.

For coupled human-environment systems, the behavioral rules of agents are usually the result of combining effects of environmental, socio-economic, and institutional factors. For example,

[*] Corresponding author at: Civil & Environmental Engineering Department, University of Michigan, 2350 Hayward St, Ann Arbor, MI, 48109, USA.
E-mail address: huya@umich.edu (Y. Hu).

http://dx.doi.org/10.1016/j.advwatres.2017.08.009
0309-1708/© 2017 Elsevier Ltd. All rights reserved.

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). Other 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). However, neither the rule-based approach nor the optimization-based approach is sufficient to capture the behavioral uncertainty arising from the bounded rationality of agents' decision-making processes. Models developed under these approaches usually do not fully reflect observed facts and phenomena, which can raise concerns when validating modeling of agents' behaviors within ABM (Elsawah et al., 2015).

However, it would be prohibitive to pinpoint the origin of agents' bounded rationality case by case and simulate them explicitly. Instead, this paper proposes an alternative approach, which presents a “gray 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 variables, and obtain the gray 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 gray box can then be fed by the data of the decision variable and its major factors to predict agents' decisions given these factors.

This paper is organized with the goal of deriving agents' behavioral rules under the impact of bounded rationality using a combined data-driven approach and domain expertise. In the next section, we first present general concepts and models necessary to introduce our methodology. Following that, we propose a methodological framework to derive agents' behavioral rules, use a case study to demonstrate the proposed framework, and present results. Finally, we conclude with our findings on the methodology and results.

2. Background: concepts and models

Agents' behavior reflects their cognitive processes of decision-making. They may be modeled 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 thorough 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., 2015b). 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 to accommodate 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), ways to measure human behaviors directly, unlike measuring physical quantities, are limited. Some aspects, for instance 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, in particular good quality behavioral data, makes 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 neglect of a confounding variable, which is an extraneous variable that correlates with other variables in a statistical model. For example, considering the DNA of two non-twin brothers, their DNA would be highly correlated, even when the DNA of non-relatives is known. However, once the DNA of the parents is known, then conditioned on the parents' DNA, the DNA of the brothers would be statistically independent. Thus, the DNA of the parents would be a confounding variable in that case. If it is not known, then a spurious causal relationship between the brothers could have been inferred. To rule out spurious relationships, this study incorporates 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' decisions and the factors. Based on the DIGs for different agents, a machine learning technique called boosted regression trees (BRT) is applied to converting the DIGs to the behavioral rules for different agents.

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). PGMs are the representations of the probabilistic relationships between 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).

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.

One major research thrust in the PGM literature is inferring the network topology – who is influencing or interacting with whom. 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 variables, dynamic Bayesian networks can be applied to finding a Bayesian network to characterize their relationships over time. Each variable 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 observations, making structure learning challenging. 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.

2.2. Granger causality

We now discuss the framework leading to the graphical model employed in this work. Granger proposed the definition of causality for a network of autoregressive time series in the 1960s: “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 information theoretic quantity known as directed information generalizes the Granger causality test and can capture such non-linear relationships to determine Granger causality (Quinn et al., 2011; Amblard and Michel, 2011).

2.3. Directed information

Given three random processes **X, Y** and **Z**, let $\mathbf{X}^{t1:\,t2}$ denote the process **X** from t_1 to t_2 inclusive and $\mathbf{X}^{t1}$ when $t_1 = t_2$. The direction information (DI) from **X** to **Y** given **Z** is defined as (Marko, 1973; Kramer, 1998)

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

the time-average expected log-likelihood ratio between two distributions of $\mathbf{Y}^{t}$, one conditioning on the past of **X**, $\mathbf{X}^{1:t\text{-}1}$, and the other not. If the past of **X** conveys no novel information for describing the future of **Y**, then the two conditional distributions will be identical, in which case the logarithm would be zero. In other terms, the present of **Y**, $\mathbf{Y}^{t}$, would be conditionally independent of the past of **X**, $\mathbf{X}^{1:t\text{-}1}$, given the past of the rest of the network, $\{\mathbf{Y}^{1:t\text{-}1}, \mathbf{Z}^{1:t\text{-}1}\}$. If the past of **X** does convey useful information, then the conditional distributions will not match. The expected log-likelihood of their ratio will be positive. The directed information measures, in bits, how well the past of **X** helps to predict the future of **Y**, even when the past of **Y** and **Z** are already known.

Directed information quantifies statistical causation in the sense of Granger causality (Quinn et al., 2015). It is well-defined for any joint distribution **P** over multiple agents. Specifically, it is not restricted to parametric models such as linear models, and it can be used for both discrete-valued and continuous-valued data. However, even when measured with directed information, latent confounding factors can lead to spurious relationships. For instance, if **X** influences **Z** and **Z** influences **Y**, then the directed information can be positive, i.e.,$\mathbf{I}(\mathbf{X} \to \mathbf{Y}) > 0$, suggesting a connection when data from **Z** is not recorded and **Z** becomes a latent confounding factor. If **Z** was observed, then the directed information would be 0, i.e., $\mathbf{I}(\mathbf{X} \to \mathbf{Y} || \mathbf{Z}) = 0$. That correctly implies that **X** has not any direct relationship with **Y**. This presents a challenge to use directed information to quantify causal relationships in real-world applications. Thus, when necessary, expert domain knowledge (i.e. human intelligence) is applied to ruling 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 that the influences found are not due to overfitting.

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 $\mathbf{X}_j$ to $\mathbf{X}_\mathbf{i}$ (for $i, j \in \{1, ..., m\}$) if $I(\mathbf{X}_j \to \mathbf{X}_i \parallel \underline{\mathbf{X}}_{\{\mathbf{1},\ldots,\mathbf{m}\}\setminus\{\mathbf{i},\mathbf{j}\}}) > 0$ (Quinn et al., 2011; Amblard and Michel, 2011). There is a directed edge from $\mathbf{X}_j$ to $\mathbf{X}_i$, if and only if knowing the past value of $\mathbf{X}_j$ helps to predict the future of $\mathbf{X}_i$, even when conditioned on the past of all other processes in the network. DIG can be used to derive the causal relationship of variables that are likely to affect agents’ behavior. Once the causal relationships between random processes are identified through the construction of DIGs, these relationships can then be translated into deterministic or probabilistic rules (i.e., agents’ behavioral rules) using various models, such as regression models. In the following, we will introduce boosted regression trees (BRT), one of the commonly used machine learning algorithms for regression analysis.

2.5. Boosted regression trees

Boosted Regression Trees (BRT) is a machine learning algorithm that uses boosting techniques to combine a large number of relatively simple tree models. BRT adaptively optimizes the predictive power of the ensemble tree models (Roe et al., 2005; Elith et al., 2008; 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 predictive performance of the current models, which is measured by the deviance between the sample data and the fitted values, namely the 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 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 (tc; 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 demonstrates the overall best predictive performance over the other methods. However, BRTs alone identify correlative relationships between variables. For the purposes of this work, we attempt to infer causal relationships that describe the behaviors of agents in an ABM. Thus, we will use BRTs along with DIG.

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.

3. Methodology

We next describe the proposed methodology. Fig. 1 outlines the steps. We define a DIG, **G**: =(**V, E**) where the set of vertices is

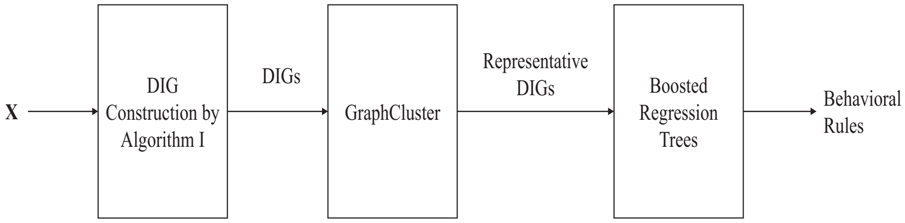


Fig. 1. Procedures to derive behavioral rules for agents. **X** represents the random processes related to agents' decision-making.

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 estimating the direct information from process $\mathbf{X}_j$ to $\mathbf{X}_i$ conditioned on the past of all other processes in the network. To estimate DI from data, we propose calculating the error variances of predictive models fit under two scenarios: 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 scenarios I and II are denoted by $\sigma_{\mathbf{I}}$ and σ_{II} respectively. The corresponding direction information from process $\mathbf{X}_j$ to $\mathbf{X}_i$ defined by Eq. (1) can be calculated using the logarithm of the ratio of error variances σ_I^2 to σ_{II}^2 (See Appendix A.1), which is then compared with the minimum description length (MDL), a model complexity penalty term to determine if the estimated influence of $\mathbf{X}_j$ on $\mathbf{X}_i$ is simply due to overfitting (Grünwald, 2007). The MDL penalty is $\mathbf{h}\log(\mathbf{n})/2\mathbf{n}$, where **h** denotes the Markov order and **n** denotes the number of samples used for mode fitting. We repeat the procedure for all the vertices in the network. As a result, we obtain the DIG as described by Algorithm 1 from Quinn et al. (2015):

Algorithm 1 Construction of the directed information graph.

```
Input: m: random processes $\{\mathbf{X}_1 \ldots \mathbf{X}_m\}$
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 $I(\mathbf{X}_j \to \mathbf{X}_i || \underline{\mathbf{X}}_{[m]\setminus\{i,j\}}) > 0$ then
            $\underline{G}(i) \leftarrow \underline{G}(i) \cup \{j\}$
        end
    end
    return $\underline{G}$
end
```

The computational complexity of Algorithm 1 is $\mathbf{O}(\mathrm{m}^2)$ if the directed information values are calculated beforehand. The algorithm can also be implemented in parallel, since each DI value can be computed separately.

As discussed above, latent confounding variables can lead to erroneous connections. To rule out the impossible directions of influence, we propose integrating expert domain knowledge with Algorithm 1. Specifically, experts will be required to confirm the plausibility of the edges inferred by the algorithm. Any edges that are deemed implausible will be removed. Especially for the scenarios with spurious confounding factors, domain knowledge can significantly improve the accuracy of the analysis. We then select the target vertices 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 a 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 (Reforgiato et al., 2008). The clustering algorithm proceeds in two phases: 1) find the highly connected substructures (i.e., the shortest path from one vertex to the other vertices in the graph) in each graph; 2) use 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 computational complexity for the agent-based modeling. Then, BRT is used to convert the DIGs for different agents that describe the causal relationships between the selected vertices and target vertices into ensemble tree models. As a result, these ensemble tree models are then used as gray boxes to represent agents' behavioral rules, and simulate/predict agents' decisions as shown by Eq. (2):

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

where **D** represents the decision variable and $\underline{V}$ is a set of variables identified by the DIG 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 left out, such as constant variables like soil types.

3.1. Case study: setup

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 a 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 provides 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 Fig. 2.

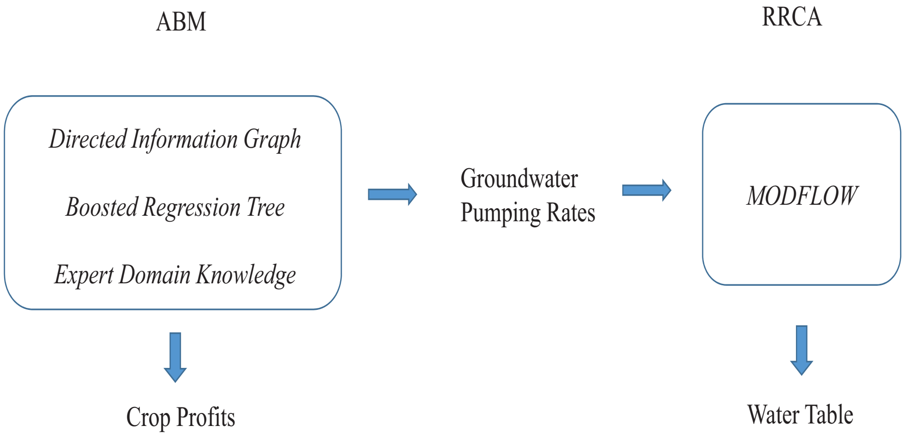


Fig. 2. Coupling of the agent-based model (ABM) with the RRCA groundwater model. The ABM is developed through the combination of the data-driven approach (i.e., DIG and BRT) with expert domain knowledge, and the RRCA model is simulated by MODFLOW-2000. The simulation period is from year 1993 to 2006.

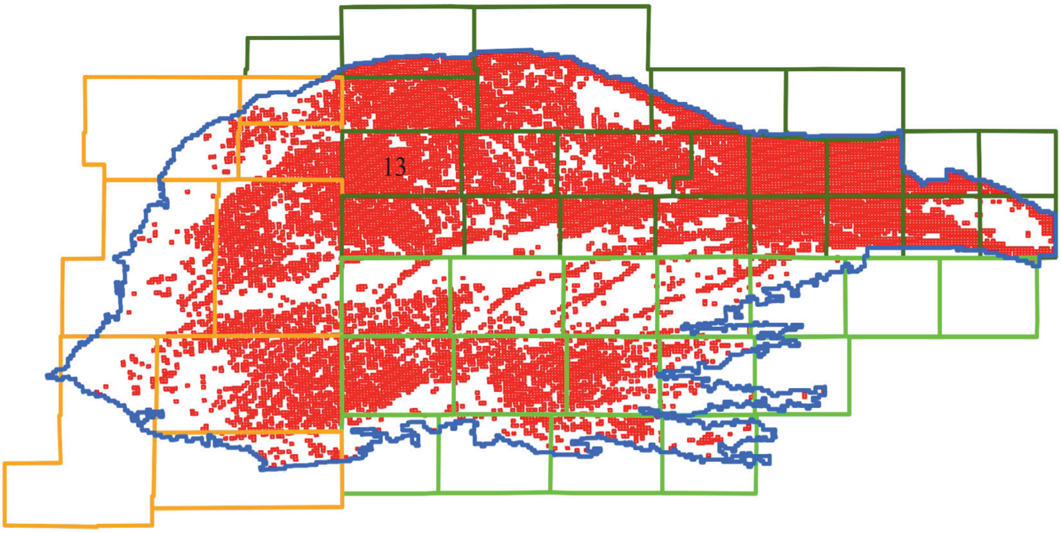


Fig. 3. The plain view of the pumping wells (red dots) and High Plains aquifer (blue line) in MODFLOW-2000 and the overlapping counties (blocks) of different states (orange: Colorado; light green: Kansas; spruce green: Nebraska; each county is treated as an super-farm agent and the number is the selected agent ID; see Hu et al., 2015a,b). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Since 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 Fig. 3, leading to 46 agents in total. Note that each farmer in the county is presumably making different decisions based on different kinds of bounded rationality. An aggregation of these individual farmers into a county level super-farm agent is not supposed to cancel out all of them, and some common ones among individual farmers should exist at the country level. In this case study, we will be focused on the behaviors of super-farm agents subject to the common kinds of bounded rationality at county level.

Table 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 crop in the study area (See Fig. 8(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 over time are reflected in the form of residuals; see Eq. (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 Fig. 3) 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 identify which variables are directly 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

Table 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 [°F]	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
Infrastrcuture	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 1000.

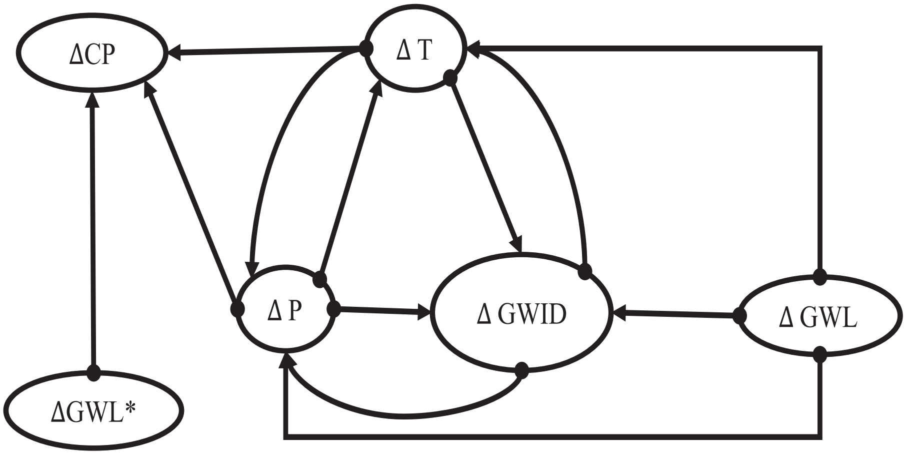


Fig. 4. Directed information graph of variables for agent 13 in Nebraska as shown in Fig. 3; the symbol Δ indicates the causal relationship is identified based on the variables with variations.

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 examine the similarities among the graphs using *GraphCluster*. As a result, some 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 form the agent-based model, which is then coupled with the RRCA model to investigate the impacts of agents' pumping behavior on the groundwater system.

4. Results and discussion

Without taking the expert domain knowledge into account, Fig. 4 shows the DIG of the variables in Table 1 using Algorithm I. For some cases, it is intuitive for domain experts 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. Fig. 4 shows that GWID causally influences 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., 2006; 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 a forcing strong enough to affect the regional temperature, cloud cover, precipitation and surface hydrology. These scientific findings help verify some parts of

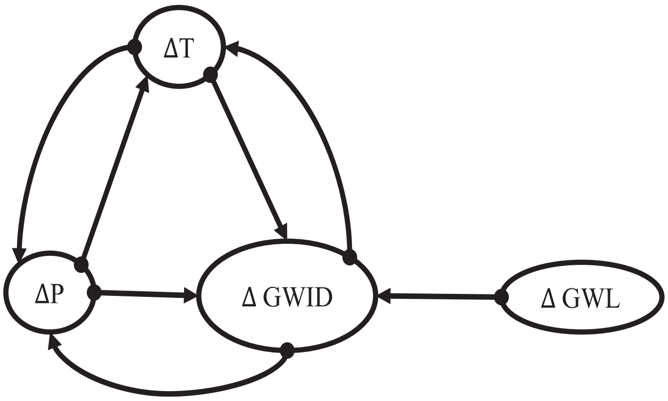


Fig. 5. Directed information graph of the variables that affect agents' behavior based on the combination of the DIG algorithm with cross-validation and the expert domain knowledge.

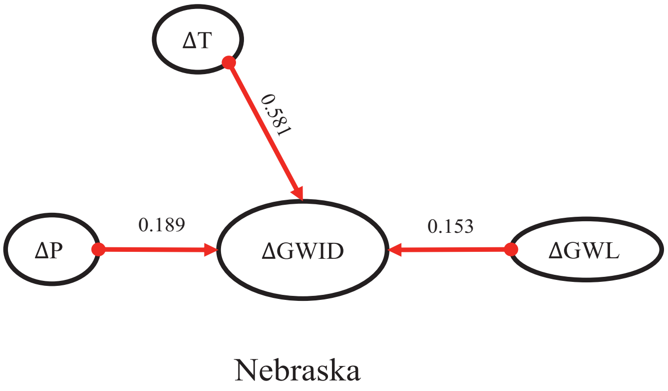


Fig. 6. Directed information graph of Chase County Agent's decision on the groundwater irrigation depth. The numbers on the edges are the measures of directed influence from one variable to the other.

the graph derived completely from the historic data using the DIG algorithm.

Fig. 4 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* were inferred as directly influencing corn price (CP), and GWL as affecting the regional mean T and P. Such spurious relationships can be ruled out with expert knowledge. For example, local T should not causally influence national CP, which should not be causally affected by local 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, Fig. 5 shows the DIG of the variables that affect the agents' behavior based both on the DIG algorithm with cross-validation and the expert domain knowledge.

Given our goal to derive the DIG that shows the causal relationships between the selected variables and the target variable (i.e., GWID), based on Fig. 5, we keep the incoming edges of GWID and the associated nodes (i.e., GWL, T and P), and truncate the remaining edges and nodes. Fig. 6 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 larger the value, the more influence one variable has on the other. 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 Fig. 7. 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 they display certain types of spatial patterns. In the following, we attempt to explain the formation of the spatial patterns.

The three representative DIGs over the study site is explained with some spatially distributed factors as shown in Fig. 8 (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 Fig. 8(a) overlap with the region circled also in blue in Fig. 8(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 to water application to satisfy the crops' 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 Fig. 8(c). Different from these three agents, agents 1 and 18 have relatively high but stable annual precipitation. As shown in Fig. 8(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 Fig. 7. For most of these agents (except agents 30, 44 and 45), their corn acreage is over 70% of their cropland area as shown in Fig. 8(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 Fig. 8(f); in particular, agents 30 and 31 (circled in purple) in Fig. 8(a) have the lowest well density over the study area as shown in Fig. 8(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' decisions on GWID (agents marked in red in Fig. 7). Referring to Fig. 8(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 relatively 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 Fig. 8(h) and shallow pumping wells are used in these areas. Thus, although there are relatively 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 Fig. 8. Agents' decision-making is complex and the three simple DIGs may not represent all agents' behavior 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.

The three representative graphs are converted to ensemble tree models using BRT. For implementation, the MATLAB function *fitensemble* is called using the least-squares boosting method *LSBoost*. 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. To avoid overfitting, 10-fold cross-validation is applied for training. Fig. 9 shows the tree No. 1 of 500 trees

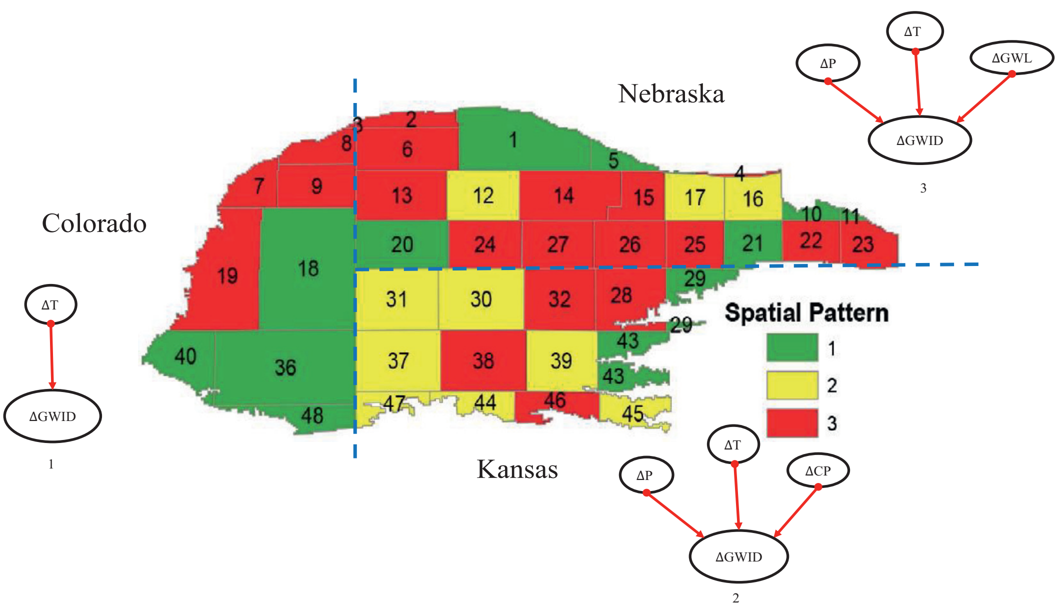
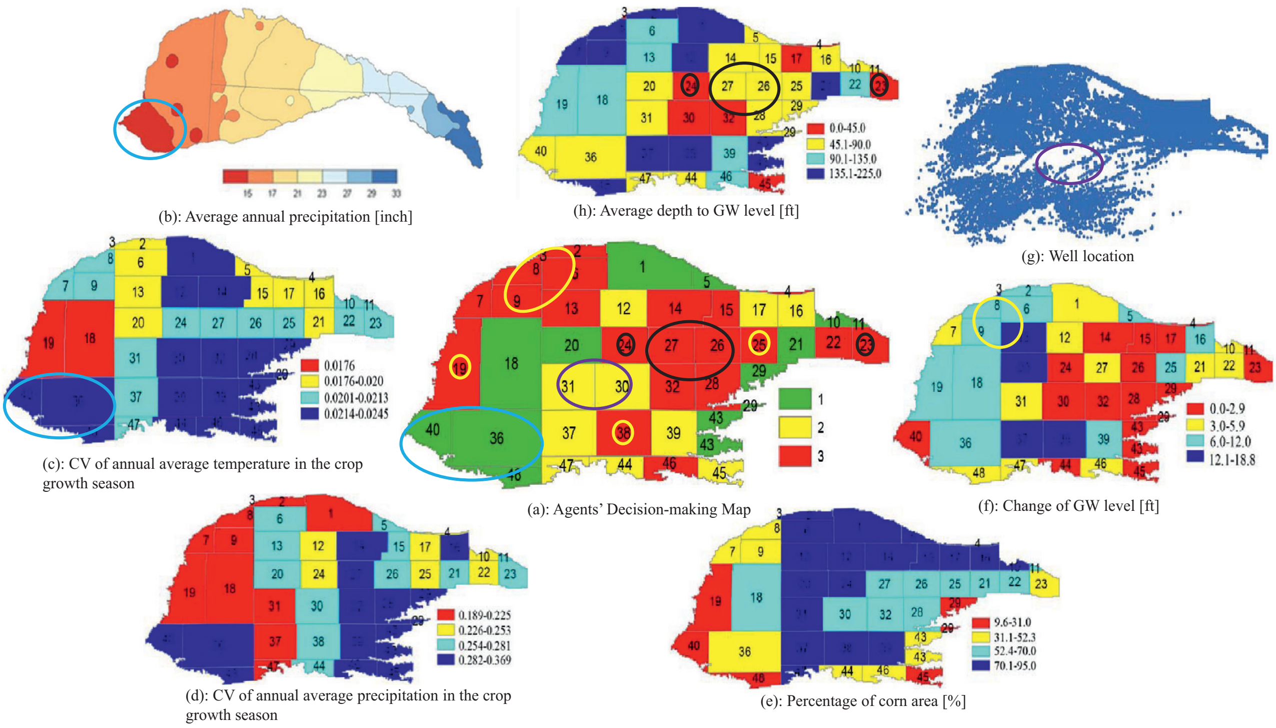


Fig. 7. Color-coded map with three directed information graphs that represent agents' decisions on the groundwater irrigation depth within the High plains aquifer HO area. The dashed lines are boundaries between different states. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



(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 [%]

Fig. 8. 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 [inches]; (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. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

for agent 24 and the ensemble tree model **M** is the linear combination of the 500 trees. Agents' monthly GWID is then computed through these tree models. 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 Fig. 9.

Given the tree model **M**, we can then compute GWID. Fig. 10 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 and 2006. Notice that we only consider GWID during the crop planting/growing season from

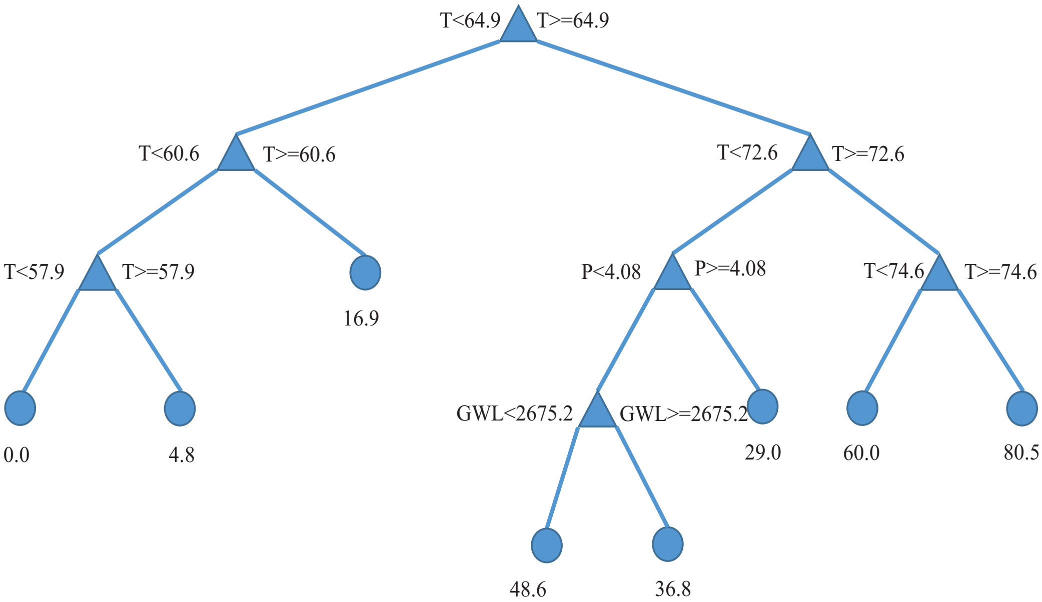


Fig. 9. Tree No. 1 of the 500 trees for agent 24; T: monthly mean temperature [°F]; GWL: groundwater level [ft]; P: monthly mean precipitation [inch]; solid blue nodes are monthly groundwater irrigation depth [mm]. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

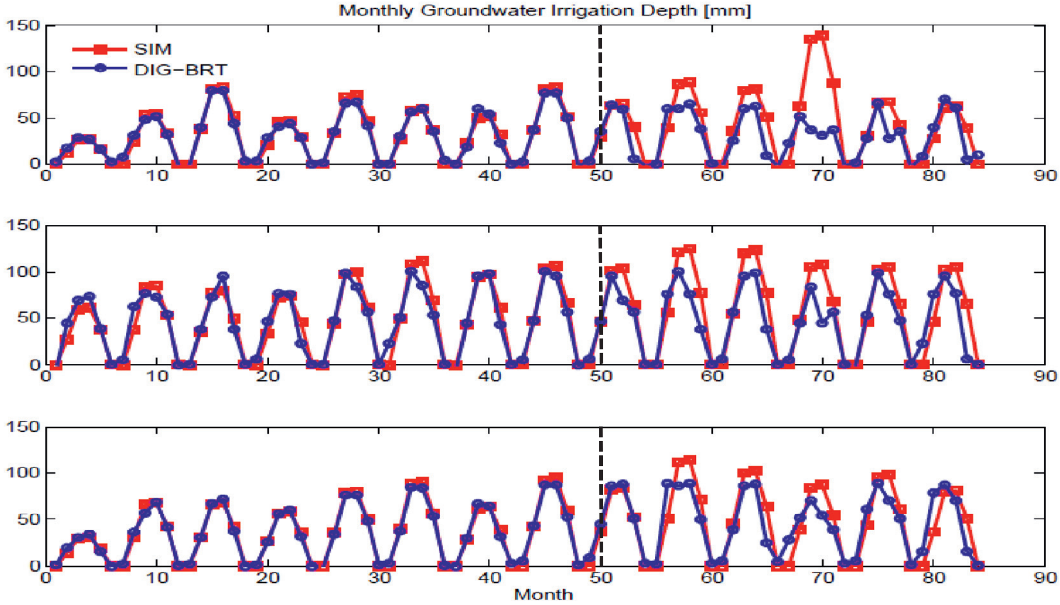


Fig. 10. Comparisons of monthly groundwater irrigation depth between the observation (red) and the model simulation using boosted regression trees model (blue) for agents 17, 18 and 24 (from top to bottom) between year 1993 and 2006. The dashed line separates the training datasets from the validation ones. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

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 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 determine these factors and to mitigate the discrepancy between the observation and the simulation.

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 Fig. 11, 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 trees (DIG-BRT) scenario.

The agents' behavioral rules derived using DIG and BRT attempt to mimic actual agents' behavioral rules. Fig. 12(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 Fig. 11. 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 remaining agents under these two scenarios, and similar results are found. We think the high goodness of fit is the result of

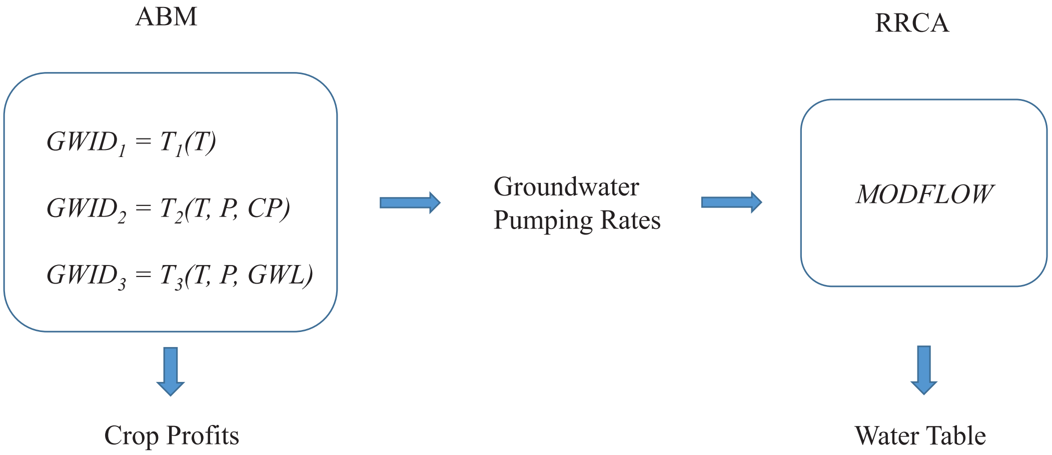


Fig. 11. Coupling of agent-based model (ABM) with the groundwater model (RRCA); three representative graphs are simulated by the corresponding BRT models denoted by T_1, T_2 and T_3. The simulation period is from year 1993 to 2006.

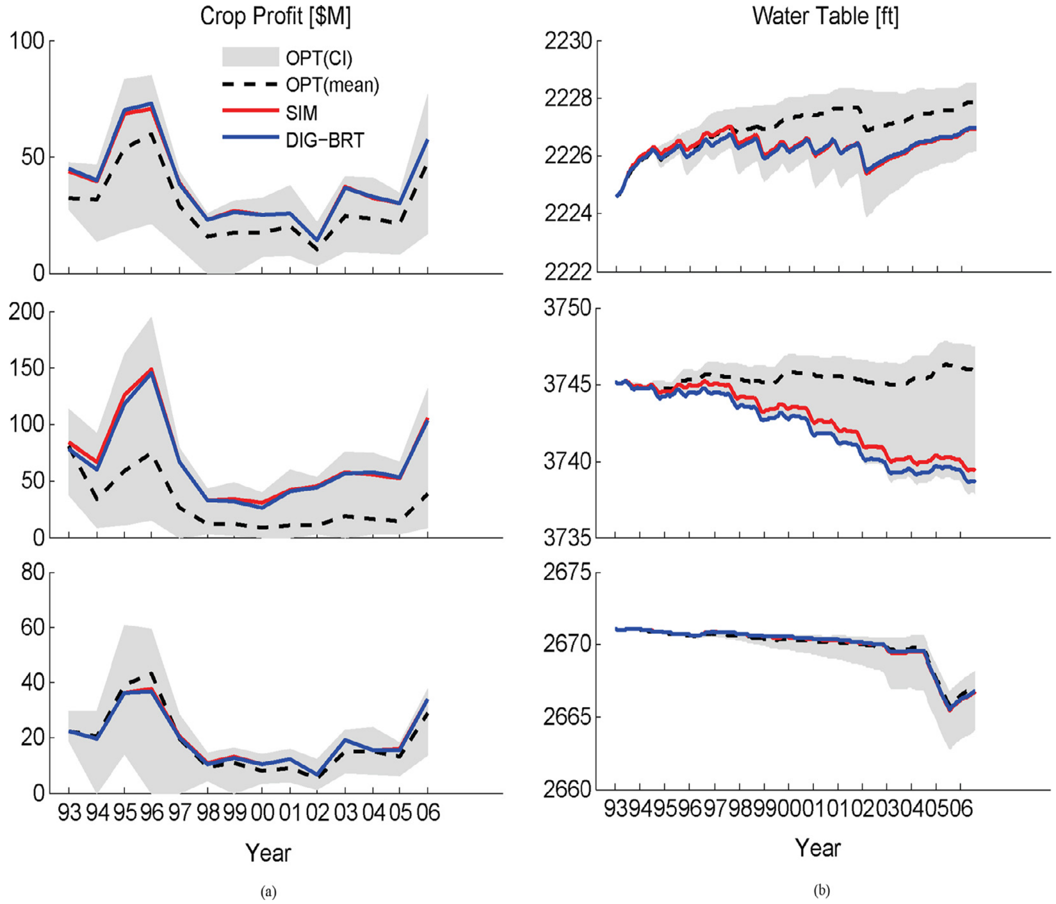


Fig. 12. Comparisons of crop profits ($M; a) and water table (ft, b) between the simulation scenario (red), the DIG-BRT scenario (blue) and the optimization scenario (the shaded area is the confidence interval and the dashed line is the mean value) for agent 17, 18 and 24 (from top to bottom). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

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.

Fig. 12(a) and (b) also 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 (see Hu et al., 2015b for more details). The shaded area indicates the confidence intervals of crop profits and water table for agents 17, 18 and 24 as the result of 1000 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 to simulate the actual agents' behavior.

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 paper, 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.

Due to missing confounding variables, insufficient sample frequency and noisy data, some of the causal relationships identified by the DIG algorithm can be spurious. In the case study, both expert domain knowledge and cross-validation are employed in the DIG algorithm to rule out spurious relationships. Through the combination of human and machine intelligence (i.e., expert knowledge, DIG and BRT), we derive the behavioral rules (i.e., via regression models) to describe agents' decision-making while accounting for their bounded rationality. In addition, it is worth mentioning that rather than limiting to regression models, various ways can be used to describe causal relationships as behavioral rules, such as probabilistic models (Ariyaratne, 2016). Future work will be conducted to investigate those alternative ways to model the causal relationships.

Note that each estimated relationship inferred statistically will improve in accuracy with more samples over the observed range of explanatory variables. However, such models might not be accurate outside the ranges observed. For instance, if all the observational data is recorded during the summer time, but the inferred relationships are applied to modeling agents' behaviors during the winter time, the results could be inaccurate. Thus, cautious steps need to be taken to examine these causal relationships.

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 a 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 an optimization-based approach (Hu et al., 2015b), 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, especially in terms of capturing the uncertainty of agents' behavior as the result of bounded rationality and mimicking their actual behavior.

Appendix A

A.1. Sufficiency of marginal normality

For special classes of distributions, the directed information

$$\mathbf{I}(\mathbf{X}\to\mathbf{Y}||\mathbf{Z})=\frac{1}{\mathbf{n}}\sum_{\mathbf{t}=1}^{\mathbf{n}}\mathbf{E}_{P_{X,Y,Z}}\left[\log\frac{\mathbf{P}_{Y_t|Y^{t-1},X^{t-1},Z^{t-1}}(\mathbf{Y}_t|\mathbf{Y}^{t-1},\mathbf{X}^{t-1},\mathbf{Z}^{t-1})}{\mathbf{P}_{Y_t|Y^{t-1},Z^{t-1}}(\mathbf{Y}_t|\mathbf{Y}^{t-1},\mathbf{Z}^{t-1})}\right], \tag{A.3}$$

simplifies to a closed, parametric formula. For instance, even though the expectation in equation A.3 is with respect to the joint distribution, $\mathbf{P}_{\mathbf{X,Y,Z}}$, if the conditional distributions $\mathbf{P}_{\mathbf{Y}_t|\mathbf{Y}^{1:t-1},\mathbf{Z}^{1:t-1}}$ and $\mathbf{P}_{\mathbf{Y}_t|\mathbf{Y}^{1:t-1},\mathbf{X}^{1:t-1},\mathbf{Z}^{1:t-1}}$ are each normal distributions with zero mean and standard deviations σ and σ' respectively, then $\mathbf{I}(\mathbf{X}\to\mathbf{Y}\parallel\mathbf{Z})=\log\sigma/\sigma'$ (Barnett et al., 2009). Note that those conditional distributions can be normally distributed without $\mathbf{P}_{\mathbf{Y}_t}$ also being normal. The derivation is included below to familiarize readers with the equation A.3:

$$\begin{aligned}
&\mathbf{I}(\mathbf{X}\to\mathbf{Y}\parallel\mathbf{Z})\\
&=\frac{1}{\mathbf{n}}\sum_{\mathbf{t}=1}^{\mathbf{n}}\mathbf{E}_{\mathbf{P}_{\mathbf{X,Y,Z}}}\left[\log\frac{\mathbf{P}_{\mathbf{Y}_t|\mathbf{Y}^{1:t-1},\mathbf{X}^{1:t-1},\mathbf{Z}^{1:t-1}}\left(\mathbf{Y}_t\middle|\mathbf{Y}^{1:t-1},\mathbf{X}^{1:t-1},\mathbf{Z}^{1:t-1}\right)}{\mathbf{P}_{\mathbf{Y}_t|\mathbf{Y}^{1:t-1},\mathbf{Z}^{1:t-1}}\left(\mathbf{Y}_t\middle|\mathbf{Y}^{1:t-1},\mathbf{Z}^{1:t-1}\right)}\right]\\
&=\frac{1}{\mathbf{n}}\sum_{\mathbf{t}=1}^{\mathbf{n}}\mathbf{E}_{\mathbf{P}_{\mathbf{X,Y,Z}}}\left[\log\mathbf{P}_{\mathbf{Y_t}|\mathbf{Y^{1:t-1}},\mathbf{X^{1:t-1}},\mathbf{Z^{1:t-1}}}\left(\mathbf{Y_t}|\mathbf{Y^{1:t-1}},\mathbf{X^{1:t-1}},\mathbf{Z^{1:t-1}}\right)\right]\\
&\quad-\frac{1}{\mathbf{n}}\sum_{\mathbf{t}=1}^{\mathbf{n}}\mathbf{E}_{\mathbf{P}_{\mathbf{X,Y,Z}}}\left[\log\mathbf{P}_{\mathbf{Y_t}|\mathbf{Y^{1:t-1}},\mathbf{Z^{1:t-1}}}\left(\mathbf{Y_t}|\mathbf{Y^{1:t-1}},\mathbf{Z^{1:t-1}}\right)\right]\\
&=\frac{1}{\mathbf{n}}\sum_{\mathbf{t}=1}^{\mathbf{n}}\left[-\mathbf{h}\left(\mathbf{Y}_t\middle|\mathbf{Y}^{1:t-1},\mathbf{X}^{1:t-1},\mathbf{Z}^{1:t-1}\right)\right]-\frac{1}{\mathbf{n}}\sum_{\mathbf{t}=1}^{\mathbf{n}}\left[-\mathbf{h}\left(\mathbf{Y}_t\middle|\mathbf{Y}^{1:t-1},\mathbf{Z}^{1:t-1}\right)\right]\\
&=\frac{1}{\mathbf{n}}\sum_{\mathbf{t}=1}^{\mathbf{n}}\left[-\frac{1}{2}\log\left((2\pi\mathbf{e})\sigma'^2\right)\right]-\frac{1}{\mathbf{n}}\sum_{\mathbf{t}=1}^{\mathbf{n}}\left[-\frac{1}{2}\log\left((2\pi\mathbf{e})\sigma^2\right)\right]\\
&=\frac{1}{2}\left[\log\sigma^2-\log\sigma'^2\right]\\
&=\log\sigma/\sigma',
\end{aligned}$$

where $\mathbf{h}(\cdot)$ denotes (differential) entropy (See Cover and Thomas, 2012, P249).

References

Aalders, I., 2008. Modeling land-use decision behavior with bayesian belief networks. Ecol. Soc. 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. J. Comput. Neurosci. 30 (1), 7–16.

Anderson, J.R., 2007. How Can the Human Mind Occur in the Physical Universe?. Oxford University Press, New York, NY.

Ariyaratne, A., 2016. Large Scale Agent-Based Modeling: Simulating Twitter Users (Master thesis). University of Maryland, College Park.

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

Barnett, L., Barrett, A.B., Seth, A.K., 2009. Granger causality and transfer entropy are equivalent for Gaussian variables. Phys. Rev. Tett. 103 (23), 238701.

Betts, A.K., 2004. Understanding hydrometeorology using global models. Bull. Am. Meteorol. Soc. 85 (11), 1673–1688.

Bonabeau, E., 2002. Agent-based modeling: methods and techniques for simulating human systems. In: Proceedings of the National Academy of Sciences of the United States of America, 99, pp. 7280–7287. http://dx.doi.org/10.1073/pnas.082080899.

Buntine, W.L., 1996. A guide to the literature on learning probabilistic networks from data. Knowl. Data Eng. IEEE Trans. 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 Resour. Manage. 25 (5), 1509–1524. http://dx.doi.org/10.1007/s11269-010-9757-y.

Caruana, R., Niculescu-Mizil, A., 2006. An empirical comparison of supervised learning algorithms. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 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. J. Geophys. Res. 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. In: Agent-Based Models of Geographical Systems. Springer, pp. 85–105.

Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. J. Animal Ecol. 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. J. Environ. Manage. 151, 500–516.

Eltahir, E.A., 1998. A soil moisture-rainfall feedback mechanism 1: theory and observations. Water Resour. Res. 34 (4), 765–776.

Eltahir, E.A., Bras, R.L., 1996. Precipitation recycling. Rev. Geophys. 34 (3), 367–378.

Granger, C.W., 1969. Investigating causal relations by econometric models and cross-spectral methods. Econometrica 424–438.

Grünwald, P.D., 2007. The Minimum Description Length Principle MIT press.

Haddeland, I., Lettenmaier, D.P., Skaugen, T., 2006. Effects of irrigation on the water and energy balances of the Colorado and Mekong river basins. J. Hydrol. 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., 2015a. Design of a web-based application of the coupled multi-agent system model and environmental model for watershed management analysis using Hadoop. Environ. Modell. Software 70, 149–162.

Hu, Y., Garcia-Cabrejo, O., Cai, X., Valocchi, A.J., DuPont, B., 2015b. Global sensitivity analysis for large-scale socio-hydrological models using hadoop. Environ. Modell. Software 73, 231–243.

Janssen, M.A., Ostrom, E., 2006. Empirically based, agent-based models. Ecol. Soc. 11 (2), 37.

Jordan, M.I., 1998. Learning in graphical models. In: 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. J. Climate 16 (3), 496–509.

Kennedy, W.G., 2012. Modelling human behaviour in agent-based models. In: Agent-Based Models of Geographical Systems. Springer, pp. 167–179.

Kramer, G., 1998. Directed information for channels with feedback. Ph.D.dissertation, Swiss Federal Institute of Technology (ETH), Zürich, Switzerland.

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. J. Hydrol. 390 (3), 222–244.

Lloyd, S.P., 1982. Least squares quantization in PCM. Inf. Theory, IEEE Trans. 28 (2), 129–137.

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

Marko, H., 1973. The bidirectional communication theory–a generalization of information theory. Commun. IEEE Trans. 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: Melton, A.W., Martin, E. (Eds.), Coding Processes in Human Memory. Wiley, New York, pp. 373–434.

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., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., 2011. Scikit-learn: machine learning in python. J. Mach. Learn. Res. 12, 2825–2830.

Pielke, R.A., 2001. Influence of the spatial distribution of vegetation and soils on the prediction of cumulus convective rainfall. Rev. Geophys. 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. J. Comput. Neurosci. 30 (1), 17–44.

Quinn, C.J., Kiyavash, N., Coleman, T.P., 2015. Directed information graphs. Inf. Theory, IEEE Trans. 61 (12), 6887–6909.

Reforgiato, D., Gutierrez, R., Shasha, D., 2008. Graphclust: a method for clustering database of graphs. J. Inf. Knowl. Manage. 7 (04), 231–241 Code available at http://www.cs.nyu.edu/cs/faculty/shasha/papers/GraphClust.html.

Roe, B.P., Yang, H.J., Zhu, J., Liu, Y., Stancu, I., McGregor, G., 2005. Boosted decision trees as an alternative to artificial neural networks for particle identification. Nucl. Instrum. Methods Phys. Res. Sect. A: Accelerators Spectrom. Detectors Associated Equip. 543 (2), 577–584.

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. J. Land Use Sci. 2 (1), 31–55.

Simon, H.A., 1996. The Sciences of the Artificial, 136. MIT press.

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

Stanilov, K., 2012. Space in agent-based models. In: Agent-Based Models of Geographical Systems. Springer, pp. 253–269.

Tesfatsion, L., 2006. Agent-based computational economics: a constructive approach to economic theory. Handbook Comput. Econ. 2, 831–880.

Ng, T.L., 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 Resour. Res. 47, W09519. http://dx.doi.org/10.1029/2011WR010399.

Urban, C., Schmidt, B., 2001. PECS-agent-based modelling of human behaviour. 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 Math. Engineers, 1 125–139.

Yang, Y.E., Cai, X., Stipanović, D.M., 2009. A decentralized optimization algorithm for multiagent system-based watershed management. Water Resour. Res. 45 (8).