Environmental Modelling & Software 73 (2015) 231–243

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

Global sensitivity analysis for large-scale socio-hydrological models using Hadoop

Yao Hu [a, *], Oscar Garcia-Cabrejo [a], Ximing Cai [a], Albert J. Valocchi [a], Benjamin DuPont [b]

[a] Ven Te Chow Hydrosystems Laboratory, Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA
[b] Nebland Software, LLC, Green Bay, WI, USA

ARTICLE INFO

Article history:
Received 23 January 2015
Received in revised form
12 August 2015
Accepted 18 August 2015
Available online 5 September 2015

Keywords:
Multi-agent system
Socio-hydrological model
Global Sensitivity Analysis
Hadoop
Polynomial Chaos Expansion

ABSTRACT

A multi-agent system (MAS) model is coupled with a physically-based groundwater model to understand the declining water table in the heavily irrigated Republican River basin. Each agent in the MAS model is associated with five behavioral parameters, and we estimate their influences on the coupled models using Global Sensitivity Analysis (GSA). This paper utilizes Hadoop-based Cloud Computing techniques and Polynomial Chaos Expansion (PCE) based variance decomposition approach for the improvement of GSA with large-scale socio-hydrological models. With the techniques, running 1000 scenarios of the coupled models can be completed within two hours with Hadoop clusters, a substantial improvement over the 42 days required to run these scenarios sequentially on a desktop machine. Based on the model results, GSA is conducted with the surrogate model derived from using PCE to measure the impacts of the spatio-temporal variations of the behavioral parameters on crop profits and the water table, identifying influential parameters.

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

* Corresponding author. 205 N. Mathews Ave., Hydrosystems Lab., 2527D, Urbana, IL 61801, USA.
E-mail address: yaohu2@illinois.edu (Y. Hu).

http://dx.doi.org/10.1016/j.envsoft.2015.08.015
1364-8152/

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 paper, 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 paper 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.

2. Background

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 Fig. 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).

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 Fig. 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 Fig. 1 and characterized by the five behavioral parameters (κ_{pr}, ν_{pr}, κ_{prep}, ν_{prep} and λ) in Table 1 (i.e., 46 agents and 230 parameters in total). For parameters κ_{pr}, ν_{pr}, κ_{prep} andν_{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 A.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 Fig. 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.

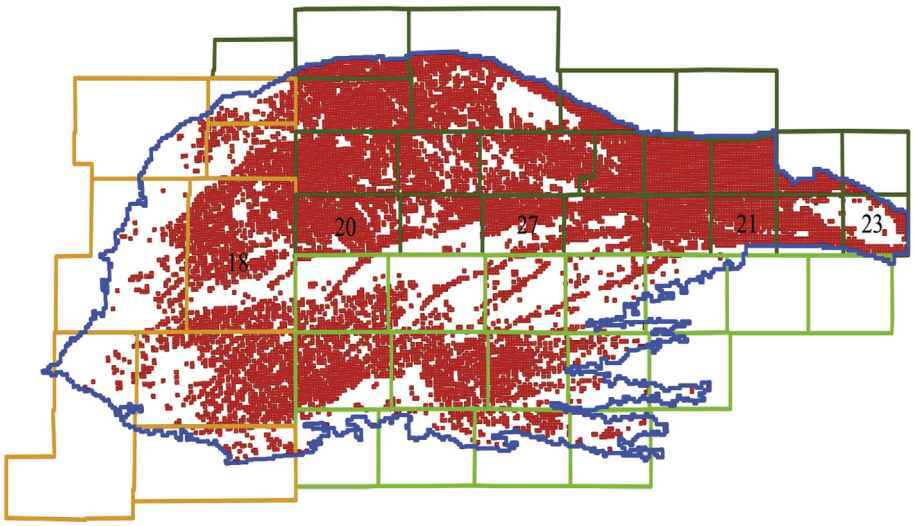


Fig. 1. The aerial view of the pumping wells (red dots) and High Plains aquifer (blue line) in MODFLOW-2000 and the overlapping counties (blocks) of different states (orange: Colorado; light green: Kansas; spruce green: Nebraska; each county is treated as an agent and the numbers are selected agent IDs). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

As a result, it can help us gain some insights into the causes and effects between farmers' pumping behaviors and groundwater decline.

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.

3.1. 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 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 $\boldsymbol{S} = \{\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(N)}\}$, where $\boldsymbol{X}^{(i)} = (\boldsymbol{X}_1^{(i)}, \ldots, \boldsymbol{X}_{230}^{(i)})$ 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 $\boldsymbol{X}^{(i)}$.

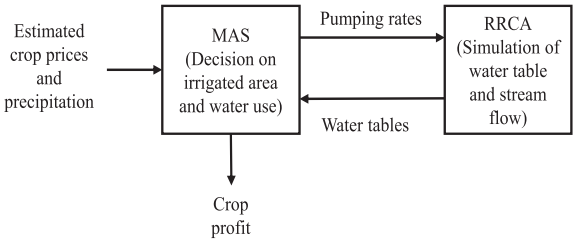


Fig. 2. Coupling of the MAS model (with an annual time interval) and RRCA model (with monthly stress period). The simulation period is from year 1993–2006.

3.2. Cloud computing with Hadoop for a large number of model runs

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 (2.4 GHz Dual Core; Hu et al., 2015). For example, if we want to run the coupled models 1000 times for sensitivity analysis (i.e., 1000 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

Table 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
$^{\kappa}pr$	[0.5,5]	Agents' beliefs in their prior knowledge of the mean of crop prices
$^{\nu}pr$	[5,50]	Agents' beliefs in their prior knowledge of the variance of crop prices
$^{\kappa}prep$	[0.5,5]	Agents' beliefs in their prior knowledge of the mean of precipitation
$^{\nu}prep$	[5,50]	Agents' beliefs in their prior knowledge of the variance of precipitation
λ	[0,20]	Agents' attitudes towards the fluctuations of crop profits

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.

3.2.1. 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 Fig. 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 $\boldsymbol{N}$ 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. Fig. 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 $\boldsymbol{N}$ 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 $\boldsymbol{N}$ simulations in the map phase is minimized as shown in the following optimization model:

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

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$\boldsymbol{N}$. 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.

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).

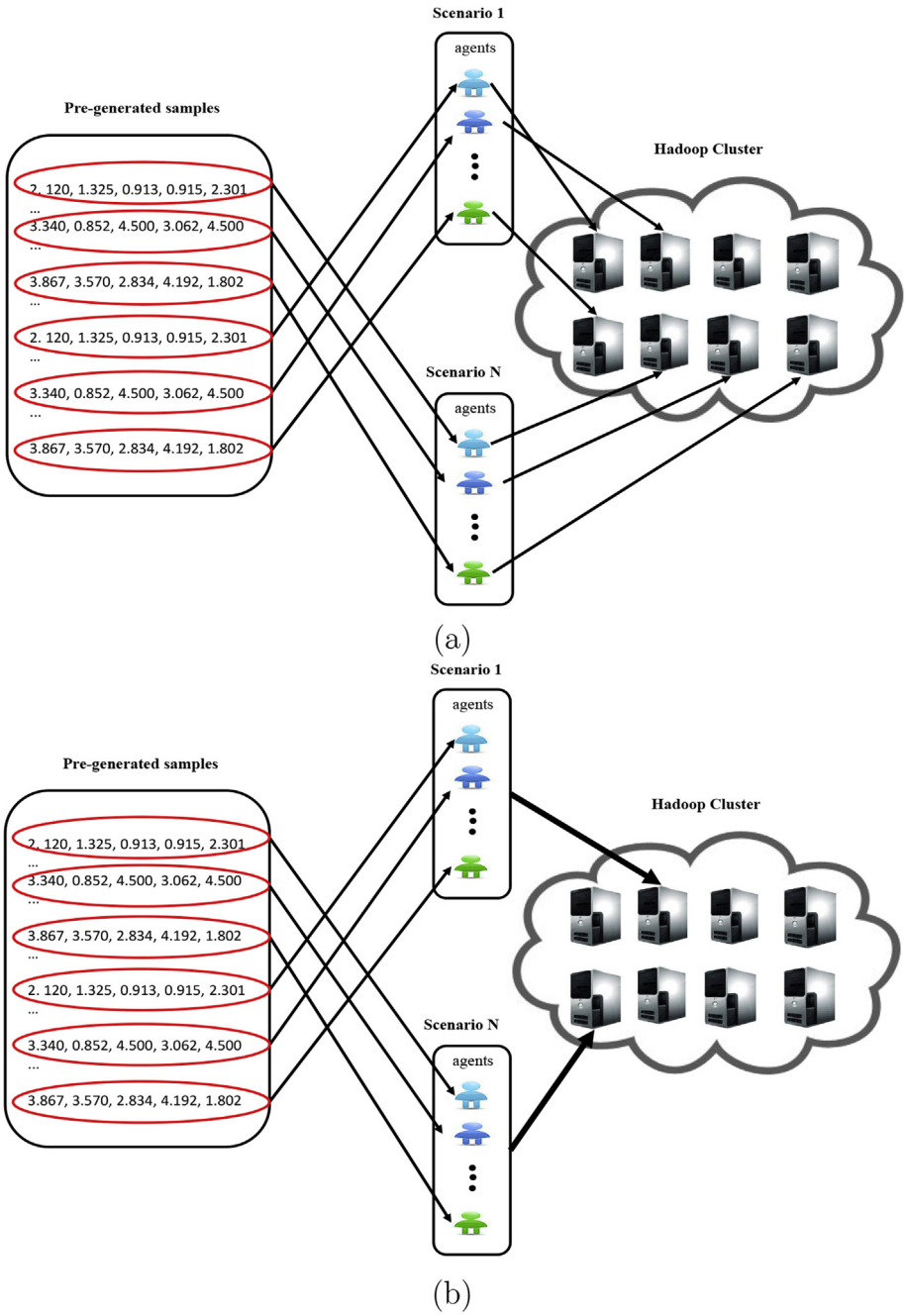


Fig. 3. (a) Schematic diagram of Approach I: agents are randomly chosen by different map tasks initiated at different machine nodes and (b) schematic diagram of Approach II: agents are lumped together and different scenarios are distributed by the MapReduce framework and executed at different machine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

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 Fig. 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 Fig. 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 Fig. 3 (b), rather than all agents from all ***N*** scenarios for every year spreading over different nodes in the first approach as shown by Fig. 3 (a).

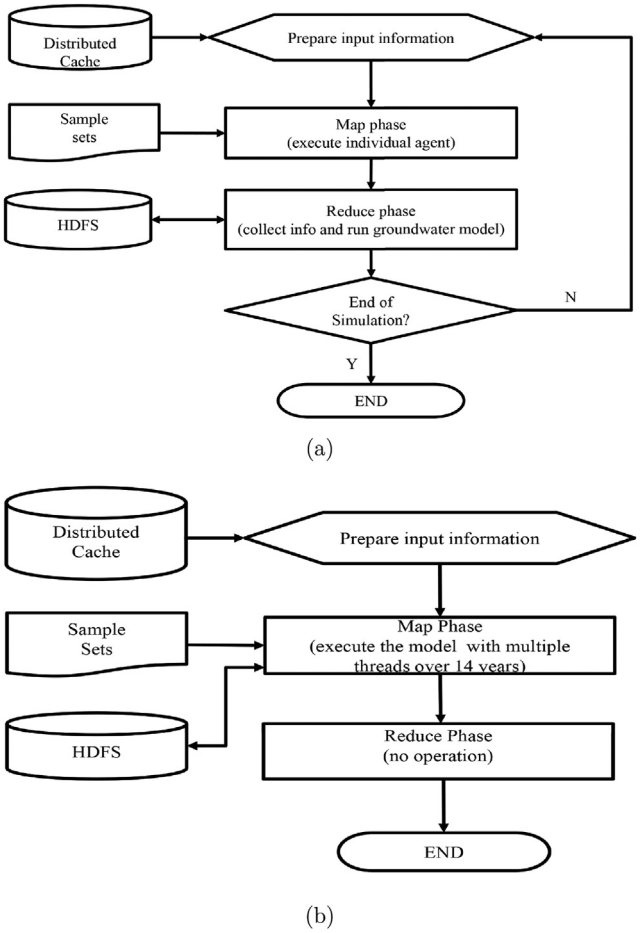


Fig. 4. (a) Approach I: Integration of the Apache Hadoop framework into the coupled MAS and RRCA model. A number of read-only files for the coupled models are stored in the distributed cache. HDFS is used to save the model outputs. Agents are spread over and executed in different map tasks initiated 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 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. Global sensitivity analysis

In the first part of the paper 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):

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

In this equation, $g_0 = E[Y]$ and $g_i(X_i) = E[Y|X_i] - g_0$ 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, $g_{i,j} = E[Y|X_i, X_j] - g_0 - g_i - g_j$ 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 Eq. (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:

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

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:

$$S1_i = \frac{V_i}{V[Y]} = \frac{Var_{X_i}\left[E_{X_{\sim i}}[Y|X_i]\right]}{Var[Y]}. \tag{4}$$

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 Eq. (3). Thus, the total effect index is given by Homma and Saltelli (1996):

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

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).

3.3.1. Polynomial Chaos Expansion

The estimation of the single and total effect sensitivity indices using Eqs. (4) and (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 Eqs. (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:

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

where β_j are a set of coefficients that define the expansion, $\xi = (\xi_1, \ldots, \xi_n)$ are independent random variables with a given PDF $\mathbf{f}_{\Xi}$, and Ψ_j are a given type of multivariate orthogonal polynomials that depends on $\mathbf{f}_{\Xi}$. For example, if ξ 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 $\Psi_j \equiv \Psi_{\boldsymbol{\alpha}} : \Psi_{\boldsymbol{\alpha}}(\xi) = \prod_{i=1}^{M} \Psi_{\alpha_i}(\xi_i)$, 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):

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

The number of terms in the PCE of Y using the multivariate orthogonal polynomial Ψ_{α} is given by $D + 1 = (M + p)!/(M!p!)$. 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

$$\overline{Y} = \beta_0 \text{ and } \sigma_Y^2 = \sum_{j=0}^{D} \beta_j^2 \left\langle \Psi_j^2 \right\rangle \tag{8}$$

where $\langle \Psi_j^2 \rangle$ is the expected value of the square of the orthogonal polynomials used in the PCE.

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 Eq. (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):

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

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

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

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

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

where the numerator is the variance of the terms of Eq. (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:

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

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.

4. Results and discussions

We assume that the five behavioral parameters are independent and uniformly distributed random variables for each agent (see Table 1). For example, Fig. 5 shows the sample distribution of the behavioral parameter sets randomly selected by agent 18 (See the agent's location in Fig. 1), and their relationships with each other.

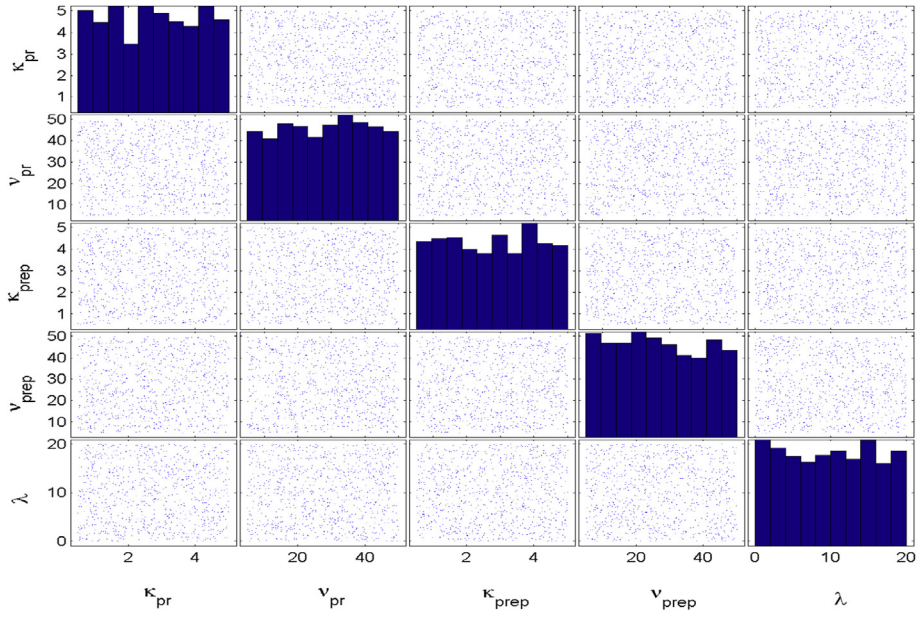


Fig. 5. Sample distribution of the behavioral parameter (κ_{pr}, ν_{pr}, κ_{prep}, ν_{prep} and λ) and their correlation with each other. The off-diagonal blocks in the Biplot show that there exist no correlation between different behavioral parameters and the diagonal blocks show that every behavioral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

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 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 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.

Table 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
Execution Time (s)	79.00	64.00	62.00	85.00	75.00	76.00
	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 (s)	74.67	66.00	59.67	90.33	77.33	74.00
Time faster (s)	2.25	2.55	2.82	1.86	2.17	2.27

In addition, according to the analysis based on Eq. (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 1000 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 234.00 s, much longer than the execution time with Approach I as shown in Table 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 1000 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 Eq. (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

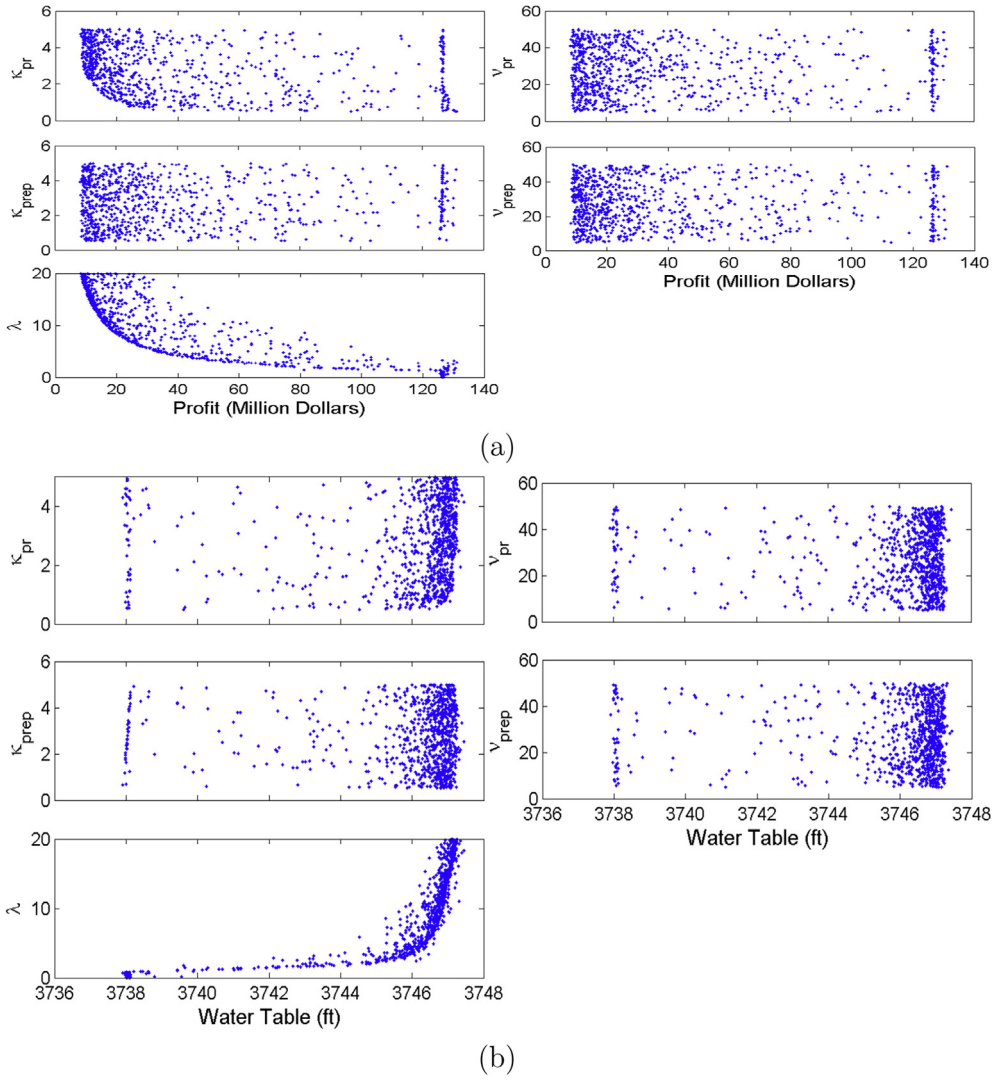
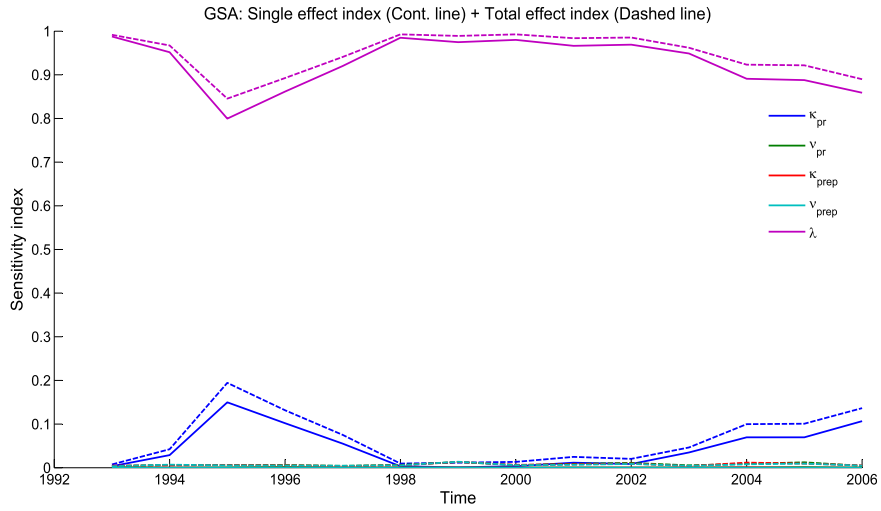


Fig. 6. (a) The relationship between behavioral parameters and crop profits for agent 18 in year 2006; (b) The relationship between the behavioral parameters and the water table in year 2006.

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 Eq. (7), that is, $p = 5$. Thus, the total number of coefficients in the PCE is 252 ($M = 5, p = 5$; 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 1000 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. Fig. 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 Fig. 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 Fig. 7(a) and (b). From the values of the single effect index (continuous line in Fig. 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



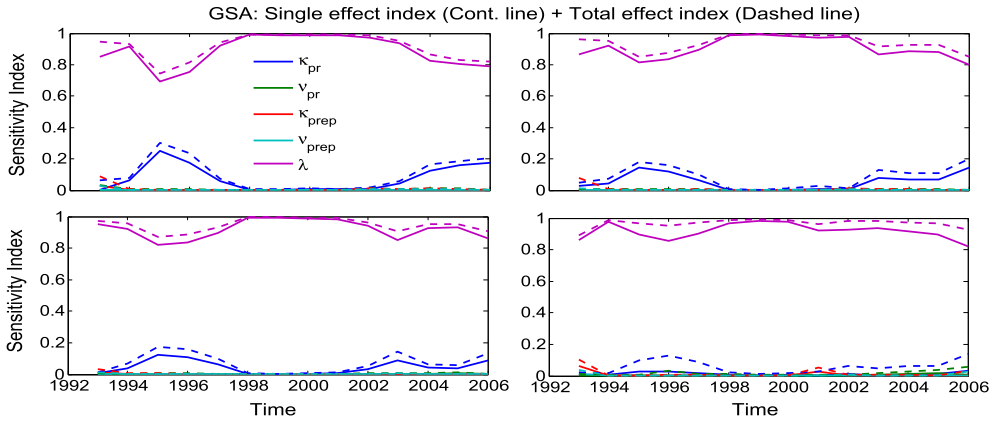
(a): GSA for crop profits

(b): GSA for the water table

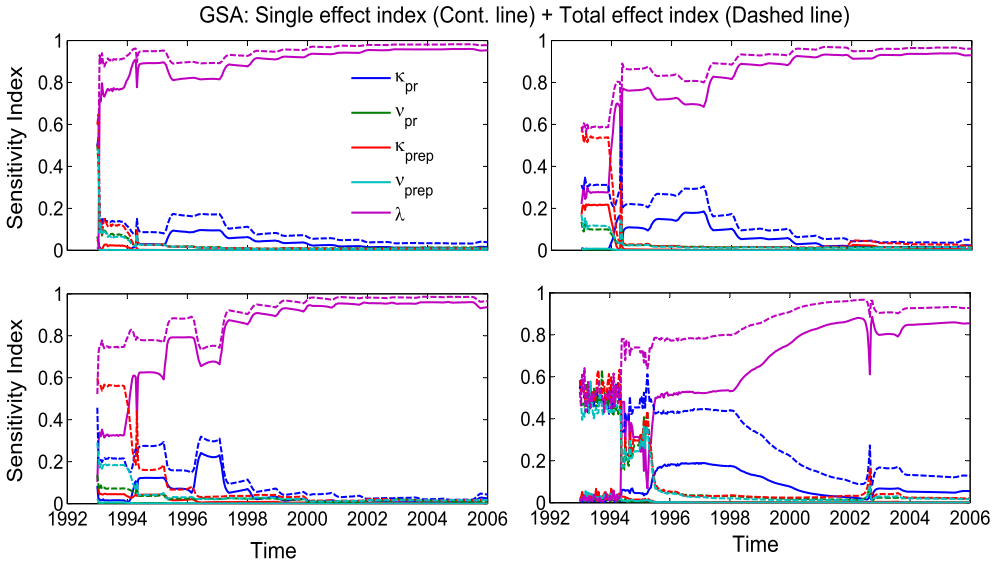
Fig. 7. (a) and (b) The temporal evolution of the sensitivity index of the behavioral parameters for agent 18 on crop profits and the water table (The continuous line and dashed line denote the single and total effect index of the behavioral parameters).

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 Fig. 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 Fig. 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 Fig. 1) on crop profits and the water table. Fig. 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 Fig. 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



(a): GSA for crop profits



(b): GSA for the water table

Fig. 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 Fig. 1) on crop profits and the water table.

the water table as the result of the dynamic interactions between parameter κ_{pr} and λ across different agents.

5. Conclusions

In this paper, 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 1000 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 1000 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. Based on the results of temporal and spatial sensitivity analysis, we are able to narrow down the focus to these two behavioral parameters while calibrating the coupled models against the real observation data.

Acknowledgments

We are graceful to the anonymous reviewers for their insightful comments and helpful suggestions. We also thank Noah Garfinkle for proofreading the article and the Department of Computer Science at University of Illinois at Urbana-Champaign for providing access to the Illinois Cloud Computing Testbed.

Appendix A

A.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:

$$p\left(D_{obs}\middle|\mu,\sigma^2\right)=\frac{1}{(2\pi)^{n/2}}\left(\sigma^2\right)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\left[n\sum_{i=1}^{n}(x_i-\overline{x})+n(\overline{x}-\mu)^2\right]\right) \tag{A.1}$$

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, 2007 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):

$$\begin{aligned}p\left(\mu,\sigma^2\right)&=NI\chi^2\left(\mu_0,\kappa_0,\nu_0,\sigma_0^2\right)\\&=N\left(\mu\middle|\mu_0,\sigma^2\middle/\kappa_0\right)\chi^{-2}\left(\sigma^2\middle|\nu_0,\sigma_0^2\right)\end{aligned} \tag{A.2}$$

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):

$$\begin{aligned}p\left(\mu,\sigma^2\middle|D_{obs}\right)&=NI\chi^2\left(\mu_n,\kappa_n,\nu_n,\sigma_n^2\right)\propto p\left(\mu,\sigma^2\right)p\left(D_{obs}\middle|\mu,\sigma^2\right)\\\mu_n&=\frac{\kappa_0\mu_0+n\overline{x}}{\kappa_n}\\\kappa_n&=\kappa_0+n\\\nu_n&=\nu_0+n\\\sigma_n^2&=\frac{1}{\nu_n}\left(\nu_0\,\sigma_0^2+\sum(x_i-\overline{x})2+\frac{n\kappa_n}{n+\kappa_n}(\mu_0-\overline{x})^2\right)\end{aligned} \tag{A.3}$$

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).

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. Comput. Geosci. 16 (3), 757–778.
Axelrod, R., 1997. Advancing the Art of Simulation in the Social Sciences. Simulating Social Phenomena. Springer, pp. 21–40.
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. In: Tools and Techniques for Social Science Simulation. Springer, pp. 243–273.
Dean, J., Ghemawat, S., 2008. Mapreduce: simplified data processing on large clusters. Commun. ACM 51 (1), 107–113. http://dx.doi.org/10.1145/1327452.1327492.
Eldred, M.S., Burkardt, J., 2009. Comparison of non-intrusive polynomial chaos and stochastic collocation methods for uncertainty quantification. 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. Reliab. Eng. Syst. Saf. 79 (2), 205–218.
Garcia-Cabrejo, O., Valocchi, A., 2014. Global sensitivity analysis for multivariate output using polynomial chaos expansion. Reliab. Eng. Syst. Saf. 126, 25–36.
Ghanem, R.G., Spanos, P.D., 1991. Stochastic Finite Elements: a Spectral Approach. Springer.
Hadoop Map/Reduce tutorial (2013). https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html.
Happe, K., 2005. Agent-based modelling and sensitivity analysis by experimental design and metamodelling: an application to modelling regional structural change. In: XIth International Congress of the European Association of Agricultural Economists.
Homma, T., Saltelli, A., 1996. Importance measures in global sensitivity analysis of nonlinear models. Reliab. Eng. Syst. Saf. 52 (1), 1–17.
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. Environ. Model. Softw. http://dx.doi.org/10.1016/j.envsoft.2015.04.011.
Huard, D., Mailhot, A., 2006. A bayesian perspective on input uncertainty in model calibration: application to hydrological model “abc”. Water Resour. Res. 42 (7),

W07416. http://dx.doi.org/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). http://dx.doi.org/10.1111/j.1745–6584.2010.00699.x.
Kelly, R.A., Jakeman, A.J., Barreteau, O., Borsuk, M.E., ElSawah, S., Hamilton, S.H., , et al.Rizzoli, A.E., 2013. Selecting among five common modelling approaches for integrated environmental assessment and management. Environ. Model. Softw. 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, third ed. Arnold Publishing.
Liebl, F., 1995. Simulation: Problemorientierte Einführung Oldenbourg.
McKusick, V., 2003. Final Report for the Special Master with Certificate of Adoption of Rrca Groundwater Model, 3605. State of Kansas v. State of Nebraska and State of Colorado, in the Supreme Court of the United States.
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. Environ. Model. Softw. 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 Resour. Res. 50 (3), 2257–2275.
Murphy, K.P., 2007. Conjugate Bayesian Analysis of the Gaussian Distribution. Technical Report. University of British Columbia.
Nielsen, M., 2009. Write Your First Map Reduce Program in 20 Minutes. Michael's Main Blog. http://michaelnielsen.org/blog/write-your-first-mapreduce-program-in-20-minutes/.
North, M.J., Macal, C.M., 2007. Managing Business Complexity: Discovering Strategic Solutions with Agent-based Modeling and Simulation. Oxford University Press.
Pappenberger, F., Beven, K.J., Ratto, M., Matgen, P., 2008. Multi-method global sensitivity analysis of flood inundation models. Adv. Water Resour. 31 (1), 1–14.
Republican River Compact Administration (RRCA), 2003. Republican River Compact Administration Ground Water Model. http://www.republicanrivercompact.org/v12p/RRCAModelDocumentation.pdf.
Saltelli, A., 2002. Making best use of model evaluations to compute sensitivity indices. Comput. Phys. Commun. 145 (2), 280–297.
Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., , et al.Tarantola, S., 2008. Global Sensitivity Analysis: the Primer. John Wiley & Sons.
Sobol, I., 1993. Sensitivity estimates for nonlinear mathematical models. Math. Model. Comput. Exp. 1, 407414.
Stocki, R., 2005. A method to improve design reliability using optimal Latin hypercube sampling. Comput. Assist. Mech. Eng. Sci. 12 (4), 393.
Storlie, C.B., Helton, J.C., 2008. Multiple predictor smoothing methods for sensitivity analysis: description of techniques. Reliab. Eng. Syst. Saf. 93 (1), 28–54.
Sudret, B., 2008. Global sensitivity analysis using polynomial chaos expansions. Reliab. Eng. Syst. Saf. 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 Res. 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 Res. 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. Comput. Geosci. 65, 84–94.
Wiener, N., 1938. The homogeneous chaos. Am. J. Math. 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 J. Sci. Comput. 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. J. Hydrology 480, 58–68.