Using Bayesian Networks for Sensitivity Analysis of Complex Biogeochemical Models

Heng Dai1,2, Xingyuan Chen1, Ming Ye3, Xuehang Song1, Glenn Hammond4, Bill Hu2, and John M. Zachara1

1Pacific Northwest National Laboratory, Richland, WA, USA, 2Institute of Groundwater and Earth Sciences, Jinan University, Guangzhou, China, 3Department of Earth, Ocean, and Atmospheric Science, Florida State University, Tallahassee, FL, USA, 4Sandia National Laboratories, Albuquerque, NM, USA

Abstract Sensitivity analysis is a vital tool in numerical modeling to identify important parameters and processes that contribute to the overall uncertainty in model outputs. We developed a new sensitivity analysis method to quantify the relative importance of uncertain model processes that contain multiple uncertain parameters. The method is based on the concepts of Bayesian networks (BNs) to account for complex hierarchical uncertainty structure of a model system. We derived a new set of sensitivity indices using the methodology of variance-based global sensitivity analysis with the Bayesian inference. The framework is capable of representing the detailed uncertainty information of a complex model system using BNs and affords flexible grouping of different uncertain inputs given their characteristics and dependency structures. We have implemented the method on a real-world biogeochemical model at the groundwater-surface water interface within the Hanford Site’s 300 Area. The uncertainty sources of the model were first grouped into forcing scenario and three different processes based on our understanding of the complex system. The sensitivity analysis results indicate that both the reactive transport and groundwater flow processes are important sources of uncertainty for carbon-consumption predictions. Within the groundwater flow process, the structure of geological formations is more important than the permeability heterogeneity within a given geological formation. Our new sensitivity analysis framework based on BNs offers substantial flexibility for investigating the importance of combinations of interacting uncertainty sources in a hierarchical order, and it is expected to be applicable to a wide range of multiphysics models for complex systems.

1. Introduction

Numerical modeling is an important tool for predicting future behaviors of complex Earth systems impacting environmental protection and water resources management (Basso et al., 2016; Dai & Samper, 2004; Tartakovsky, 2013). Predictive uncertainty is inevitable in Earth system models because of the complexity of natural system processes, physical and chemical heterogeneity in the natural environment, and limited data availability for characterizing the system properties or validating the models (Neuman, 2003; Refsgaard et al., 2012; Ye et al., 2004, 2005, 2008). To effectively and efficiently reduce predictive uncertainty with limited resources, sensitivity analysis is an essential step to rank the importance of different uncertainty sources that contribute to overall predictive uncertainty (Dai, Chen, et al., 2017; Razavi & Gupta, 2015; Wainwright et al., 2014).

Compared to local methods, global sensitivity analysis has become more popular because it can rank the importance of uncertain model inputs while considering their whole range of values and interactions (Herman et al., 2013; van Griensven et al., 2006). Conventional global sensitivity analysis methods focus on the importance of model parameters while neglecting contributions of model structures, and they do not consider the dependence structures among different uncertain inputs in a complex model system (Chu-Agor et al., 2011; Saltelli, 2000; Song et al., 2015). The single-model, parameter-oriented sensitivity analysis is insufficient for identifying dominant model processes, each of which usually consists of multiple parameters (Clark et al., 2015, 2016). Process-oriented sensitivity analysis has gained increasing attention for improving hydrological models and beyond (Dai, Chen et al., 2017, Dai, Ye et al., 2017; Grayson & Bloschl, 2000; Sivakumar, 2004, 2008).
Dai and Ye (2015) developed a multilayer hierarchical sensitivity analysis framework based on the variance-based global sensitivity analysis method (Saltelli et al., 1999; Saltelli, 1995). Dai, Chen, et al. (2017) and Dai et al. (2018) further developed a sensitivity framework capable of quantifying relative contribution of uncertainty from multiple, spatially distributed model inputs with relatively low computational cost. However, these frameworks are constrained by a simple three-layer structure of uncertainty: model parameters, model structures, and forcing scenarios, which may not be sufficient for describing a large number of uncertainty sources involved in multiphysics environmental modeling nor can they adequately characterize the complex relationships between various uncertainty sources. It is also not possible to explore the contribution of uncertainty from a group of factors across the different layers. To overcome these weaknesses, this work aims to develop a more flexible sensitivity analysis framework beyond the three-layer structure using the Bayesian network (BN; Pearl, 2014).

Also known as belief networks, BNs have increasingly gained attention across a number of scientific frontiers (i.e., artificial intelligence and machine learning) in the last few decades (Darwiche, 2009). BNs consist of a directed graphical structure and a probabilistic relationship among a set of nodes (i.e., variables; Heckerman, 1997; Velikova et al., 2014). The causal or dependence relationships are explicitly represented in the graphical structure, which allows a complex causal chain to be factored into a series of conditional relationships that are essential for deriving the joint probability distribution of a given set of variables. Thus, the BN is a powerful tool in representing a complex modeling system with various uncertainty sources and their dependence structure (Aguilera et al., 2011; Darwiche, 2009).

We propose to integrate the BN representation and inference with the variance-based method to form a new sensitivity analysis framework. In this framework, various uncertainty sources of a modeling system are represented as different nodes in the BN based on their characteristics and correlations. These uncertainty sources can be flexibly combined and analyzed together based on their causal relationships through the BN. A new sensitivity index can be derived for a combination of uncertain factors, which could be grouped in different ways depending on the questions to be answered, by combining the variance-based sensitivity analysis principle with partial variances obtained from the BN inference (explained in section 2). This index can then be used to assess their relative contribution to overall uncertainty compared to other combinations of uncertain factors.

For the purposes of methodology evaluation and demonstration, we implement the newly developed sensitivity analysis framework into a two-dimensional (2-D) groundwater reactive transport model, built based on data from the Hanford Site’s 300 Area, a U.S. Department of Energy (DOE) site (Song et al., 2018). The model contains multiple uncertain factors that are flexibly grouped into different sets of model forcing, processes, or subprocesses based on their characteristics and modeling structure. A series of sensitivity analyses is performed to answer two sequential questions: (1) Which process contributes the most to the total uncertainty of total subsurface organic carbon consumption? Then (2) which component is the most influential uncertainty contributor to the process that is identified in the first analysis? Although our test case is for groundwater reactive transport, our methodology is theoretically rigorous and general. Thus, it can be applied to a wide range of hydrologic and environmental modeling that improves our understanding of the Earth system.

2. Methodology

We start with a brief introduction of a BN and its use for groundwater biogeochemical reactive transport modeling in section 2.1. The BN-based sensitivity indices are derived in section 2.2 to identify the relative importance of each process that is involved in the groundwater reactive transport modeling.

2.1. BN for Groundwater Reactive Transport Modeling

A BN is an acyclic directed graph, with its nodes representing the stochastic variables and the directional edges describing the conditional dependence or kinship relationships among the nodes (Aguilera et al., 2013). For example, a directional edge from X to Y establishes parent-child relation between X and Y, that is, X is the parent node of Y, and Y is the child node of X. A variable in the BN is described through its conditional probability distribution or conditional probability table given its parents in the graph.
Distinct uncertainty sources are represented by different variables of the BN, and they are characterized by their continuous or discrete probability distributions. The uncertainties are then propagated through the network to the model outputs through the Bayesian inference. Notably, there also is a special type of node in the BN called “deterministic node,” whose values are directly calculated by those of their parents through deterministic equations (e.g., process-based governing equations) instead of being linked through conditional probabilities (Kjaerulff & Madsen, 2008). Based on the conditional dependence or independence contained in the acyclic directed graph, the full joint distribution of the BN can be computed as the product of conditional probabilities of all the nodes given their parents:

\[
p(X_1, X_2, ..., X_N) = \prod_{i=1}^{N} p(X_i|\text{parents}(X_i)),
\]

where \( p \) represents a probability distribution, \( X_i \) is an individual node (variable) in the BN, and \( p(X_i|\text{parents}(X_i)) \) represents the conditional probability of \( X_i \) given its parents. The full joint probability offers the basis to derive conditional probabilities of a group of variables given another group of variables, enabling the flexible decomposition of uncertainties into various sources as discussed in section 2.2.

For a general groundwater biogeochemical reactive transport modeling system, a BN could be built to represent its primary components (shown in Figure 1)—climate scenario, groundwater flow process, heat transport process, and reactive transport process. The climate scenario provides forcing to various processes in the model, including precipitation/snow, temperature, river stage, and groundwater levels. The groundwater flow, heat transport, and reactive transport processes all contain similar essential elements: initial
conditions, boundary conditions, governing equations, and parameters. The velocity field output from the flow process is used as input to the heat transport and reactive transported processes, and the temperature field output from the heat transport process changes the biochemical reaction rates associated with the reaction network informed by the microbial genomic information. The parameters for the groundwater flow, permeability field, might be dependent on both the geological structure of various formations and within-formation heterogeneity. These various model components are represented as different BN nodes in Figure 1, combined into four primary groups, that is, climate scenario, groundwater flow process, heat transport process, and reactive transport process. All the model outputs are characterized as deterministic nodes (in ovals in Figure 1) because they are governed by physical laws given known inputs. Each directional edge in the BN represents the causal dependence between two variables, with the deterministic linkages through physical laws represented using bold arrows, distinguishing from those linkages through conditional probabilities. This graphical representation of the model system demonstrates all the causal relationships between the system behaviors (i.e., model outputs) and the various uncertain factors, which greatly facilitates not only the understanding of a complex system but also communicating the scientific understanding to various stakeholders. The model input variables also are grouped by the different physical and biogeochemical processes in the model system, that is, driving force or climate scenario, groundwater flow process, heat transport process, biogeochemical reaction, and transport process, which are represented by boxes with distinct colors in Figure 1.

2.2. BN-Based Sensitivity Analysis

The core of sensitivity analysis used in this study lies in decomposing the total variance of a model output \( \Delta = f(\varnothing_1, \ldots, \varnothing_s) \) (\( f \) denotes a model and \( \varnothing_i \) a model input) into its dependent inputs (Saltelli et al., 1998, 1999, 2010; Sobol’, 1993) as

\[
\text{Var}(\Delta) = \text{Var}_S \big( E_{\varnothing_L}(\Delta|\varnothing_L) \big) + E_{\varnothing_L} \big( \text{Var}_S(\Delta|\varnothing_L) \big),
\]

where \( \text{Var} \) and \( E \) are the notations for variance and expectation calculations, respectively, and \( \varnothing_L \) represents the model inputs except the input \( \varnothing_i \). The first-order sensitivity index is defined as \( S_i = \frac{\text{Var}_S \big( E_{\varnothing_L}(\Delta|\varnothing_L) \big)}{\text{Var}(\Delta)} \), which quantifies the percentage of output uncertainty contributed by \( \varnothing_i \) and measures its relative importance compared to other uncertain inputs. The variance decomposition technique has been recursively applied by Dai, Chen, et al. (2017) to a three-layer hierarchical uncertainty framework, from which a new set of sensitivity indices is defined to quantify the uncertainty contributions from three groups of uncertain inputs: scenario, model structure, and parameters. However, the principle of variance decomposition can be applied to a hierarchical structure with more than three layers following the recursive manner. BN has a natural hierarchical structure in linking nodes through edges, which offers great flexibility to group various uncertain variables in the network and construct a dependence structure that allows recursive variance decomposition.

Using the BN shown in Figure 1 as an example, all the model inputs for generating the model output of solute concentration are combined into four groups: the climate scenario \( (P_S) \), groundwater-flow process \( (P_F) \), heat transport process \( (P_H) \), and reactive transport process \( (P_R) \). The reactive transport process depends on the groundwater flow and heat transport processes, which are both influenced by the climate scenario through boundary and initial conditions. The hierarchical structure of these four uncertainty groups in the BN enables the similar recursive decomposition of the total variance in solute concentration into individual contributors following Dai, Chen, et al. (2017). The total variance in model output \( \Delta \) can be first decomposed as

\[
\text{Var}(\Delta) = \text{Var}_{P_S}E_{P_F,P_H,P_R}(\Delta|P_S) + E_{P_S} \text{Var}_{P_F,P_H,P_R}(\Delta|P_S),
\]

where the first term on the right-hand side of equation (3) is the variance contributed from the climate scenario \( (P_S) \), and the second term is the variance caused by all other components. The variance calculation inside the second term can be further decomposed to contribution from the groundwater flow process and other groups excluding climate scenarios, which leads to
where the first term on the right-hand side of this equation represents the variance contributed from the groundwater flow process \((P_\text{F})\). Similarly, the second term on the right can be further decomposed based on the heat transport process which is dependent on both climate scenario and groundwater flow transport process:

\[
E_{P_\text{F}}E_{P_\text{H}}E_{P_\text{F}} E_{P_\text{H}}(\Delta P_S, P_F, P_H) = E_{P_\text{F}}E_{P_\text{H}}(\Delta P_S, P_F, P_H) + E_{P_\text{F}}E_{P_\text{H}}(\Delta P_S, P_F, P_H),
\]

(5)

where the two terms on the right-hand side of equation (5) represent the variances contributed from the heat transport process \((P_{\text{th}})\) and reactive transport process \((P_R)\), respectively. Substituting equations (4) and (5) into equation (3) and applying the law of total expectation, the final decomposition becomes

\[
\text{Var}(\Delta) = Var_{P_\text{F}, P_\text{H}, P_\text{F}}(\Delta P_S, P_F, P_H) + E_{P_\text{F}}E_{P_\text{H}}(\Delta P_S, P_F, P_H) + E_{P_\text{F}}E_{P_\text{H}}(\Delta P_S, P_F, P_H) = Var(P_S) + Var(P_F) + Var(P_{\text{th}}) + Var(P_R),
\]

where the four variances represent the contributions from the climate scenario \((P_S)\), groundwater flow process \((P_F)\), heat transport process \((P_{\text{th}})\), and reactive transport process \((P_R)\), respectively. Therefore, a set of sensitivity indices can be defined for each group as

\[
\begin{align*}
S_S &= \frac{Var_{P_\text{F}, P_\text{H}, P_\text{F}}(\Delta P_S, P_F, P_H)}{Var(\Delta)} = \frac{Var(P_S)}{Var(\Delta)} \\
S_F &= \frac{E_{P_\text{F}}E_{P_\text{H}}(\Delta P_S, P_F, P_H)}{Var(\Delta)} = \frac{Var(P_F)}{Var(\Delta)} \\
S_{\text{H}} &= \frac{E_{P_\text{F}}E_{P_\text{H}}(\Delta P_S, P_F, P_H)}{Var(\Delta)} = \frac{Var(P_{\text{th}})}{Var(\Delta)} \\
S_R &= \frac{E_{P_\text{F}}E_{P_\text{H}}(\Delta P_S, P_F, P_H)}{Var(\Delta)} = \frac{Var(P_R)}{Var(\Delta)}.
\end{align*}
\]

(7)

Note that these derived sensitivity indices are not the conventional first-order sensitivity indices or the total effects indices, and they sum to 1. After the sensitivity indices are evaluated, the relative importance of the four groups can be assessed. It also should be noted that although the sensitivity indices are derived for the four processes, similar indices can be derived for other combination of nodes, depending on specific interests. This is the strength of the BN-based sensitivity analysis, in that the BN (after separating model components into the BN nodes) enables us to combine the uncertain scenarios, models, and parameters into one process module.

The expectations and variances involved in partial or total variance calculations in equations (6) and (7) are obtained from the distributions assigned based on prior information or through BN inference based on equation (1) (Pearl, 2014). Using the scenario variance \(\text{Var}(P_S)\) as an example, it can be first written in the form of expectations based on the definition of variance:

\[
\text{Var}(P_S) = E_{P_\text{F}, P_\text{H}}(E_{P_\text{F}}E_{P_\text{H}}(\Delta P_S, P_F, P_H))^2 - (E_{P_\text{F}}E_{P_\text{H}}E_{P_\text{F}}(\Delta P_S, P_F, P_H))^2.
\]

(8)
After applying the definition of expectation and the law of total expectation, equation (8) can be rewritten as

\[
\text{Var}(P_S) = \sum_{P_S} p(P_S) \left( \sum_{P_F} p(P_F|P_{H}, P_R) (\Delta|P_S, P_F, P_{H}, P_R) \right)^2
\]

(9)

\[
- \sum_{P_S} p(P_S) \left( \sum_{P_F} p(P_F|P_{H}, P_R) (\Delta|P_S, P_F, P_{H}, P_R) \right)^2
\]

\[
+ \sum_{P_S} p(P_S) \left( \sum_{P_F} p(P_F|P_{H}, P_R) (\Delta|P_S, P_F, P_{H}, P_R) \right)^2
\]

\[
- \left( \sum_{P_S, P_F, P_{H}, P_R} p(P_S, P_F, P_{H}, P_R) (\Delta|P_S, P_F, P_{H}, P_R) \right)^2.
\]

Since there is no parent node existing for the nodes of process \(P_S\), and they are independent from each other, the probability \(p(P_S)\) can be rewritten as

\[
p(P_S) = \prod_{X_S \rightarrow P_S} p(X_S),
\]

(10)

where \(X_S\) represents the variables contained in the corresponding process of climate scenario \(P_S\). The probability \(p(P_S, P_F, P_{H}, P_R)\) can be estimated by explicitly expressing the full joint probability using the chain rule of conditional probability product in the BN of equation (1), and the probability \(p(P_F, P_{H}, P_R)\) can be estimated via marginalization and the equation (9) becomes

\[
\text{Var}(P_S) = \sum_{P_S} \left( \prod_{X_S \rightarrow P_S} p(X_S) \right)
\]

(11)

\[
\times \left( \sum_{P_S, P_F, P_{H}, P_R} \left( \prod_{P_S} p(X|\text{parent}(X)) (\Delta|P_F, P_{H}, P_R, P_S) \right) \right)^2
\]

\[
- \left( \sum_{P_S, P_F, P_{H}, P_R} \left( \prod_{P_S} p(X|\text{parent}(X)) (\Delta|P_F, P_{H}, P_R, P_S) \right) \right)^2.
\]

where \(X\) represents the variables contained in the BN, and parent(X) is the parent variables of \(X\). This equation affords the ability to flexibly access information from multiple grouped uncertain variables in the BN. The flexible structure of BN also allows us to further investigate or quantify the uncertainty of individual components or their combinations within the predefined process or groups of variables. For example, the groundwater flow process \((P_F)\) in Figure 1 is an aggregate group of variables that contains geologic module, permeability module, governing equations, boundary conditions, and initial conditions. If we are interested in the fraction of uncertainties contributed by the geological module \((P_G)\), including alternatives of conceptualization distribution of different geologic formations and within formation heterogeneity) and the permeability module \((P_P)\), including multiple plausible within formation permeability fields, a new set of sensitivity indices can be defined. Since the permeability module is dependent on geological module, the partial variance in equation (6) caused by flow transport process components can be decomposed further as

\[
E_{P_F} \text{Var}_{P_F, P_{H}, P_R} (\Delta|P_F, P_S) = E_{P_F} \text{Var}_{P_G, P_P, P_{H}, P_R} (\Delta|P_F, P_G, P_P) + E_{P_F} E_{P_P} \text{Var}_{P_{H}, P_R} (\Delta|P_F, P_P, P_{H}, P_R)
\]

(12)

\[
+ \text{Var}(P_G) + \text{Var}(P_P).
\]

Therefore, two subprocess sensitivity indices \(S_G\) and \(S_P\) can be defined as

\[
S_G = \frac{\text{Var}(P_G)}{\text{Var}(\Delta)} \quad \text{and} \quad S_P = \frac{\text{Var}(P_P)}{\text{Var}(\Delta)}.
\]

(13)

This example illustrates the flexibility of BN-based sensitivity analysis in answering a hierarchical list of questions that are common in a multiscale complex system.
3. Implementation of Groundwater Biogeochemical Reactive Transport

We implemented the new BN-based sensitivity analysis framework to a groundwater biogeochemical reactive transport model developed for the DOE Hanford Site's 300 Area. The study site is located in southeastern Washington State on the western shore of the Columbia River (Figure 2a), which has been known as a legacy nuclear waste site. There has been numerous monitoring and characterization effort at the 300 Area to understand a persisting uranium plume within the Columbia river corridor (e.g., Chen et al., 2012, 2013; Johnson et al., 2015; Murakami et al., 2010; Murray et al., 2013; Zachara et al., 2016). The unconfined aquifer in this area consists of three distinct geological formations as shown in Figure 2b: highly permeable Hanford formation (which is mainly coarse river cobble); less permeable alluvium formation; and the least permeable Ringold Formation, which is formed from silty sand (Chen et al., 2012, 2013). The river stage in the Columbia River is highly dynamic due to the upstream dam operations (Song et al., 2018). Geochemical and biogeochemical modeling for this system is subject to multiple uncertain factors, such as the permeability field controlled by geological formations and within-formation heterogeneity, river stage and river temperature dynamics under future climate scenarios, reaction network, and associated rates impacted by the abiotic environmental conditions as well as the microbial community that are present in the system (Graham et al., 2017; Stegen et al., 2016, 2018). Therefore, sensitivity analysis is an essential step for better understanding the model system and implementing efficient strategies to reduce the uncertainty in model predictions.

3.1. Model Setup

A 2-D groundwater reactive transport model has recently been built using site-specific data to study the coupled hydrologic, thermal, and biogeochemical processes in the hyporheic corridor of the Hanford 300

![Figure 2. (a) Location map of the Department of Energy Hanford Site. The red line is the profile of the 2-D model between monitoring well W2-3 and river gauge RG3. (b) Spatial distribution of Hanford formation, Ringold formation, and alluvium layer beneath the Columbia River.](image-url)
Area (Song et al., 2018). The numerical model was built using PFLOTRAN, a massively parallel three-dimensional flow and reactive transport code (Hammond et al., 2014). The model domain is 143.2 m long in the horizontal direction and 20 m deep in the vertical direction. The domain was discretized into approximately 200,000 nonuniform grid cells, with finer grids used in the area closer to the riverbed to better capture the exchange dynamics and the associated biogeochemical processes at the groundwater-surface water interface. The model domain extended from the Columbia River on the right to an inland well on the left as shown in Figure 2b. A transient hydraulic pressure head boundary and a Dirichlet temperature boundary were used on both lateral boundaries using the monitoring data obtained from a piezometer in the river and a groundwater monitoring well at the 300 Area. No-flow and no-heat transfer boundary conditions were used at the top and bottom boundaries. Numerical tracers were placed at the right boundary to track the intrusion of river water in the model domain. The model simulation was conducted for a length of 12 weeks corresponding to one out of six climate scenarios selected from each year of 2010 to 2015 (described in section 3.2). The initial hydraulic head field under each scenario was set to be the average hydraulic heads between the inland and river boundaries at the starting time and the temperature field initialized using initial temperature at the inland boundary.

The biogeochemical reaction network and the reaction rates in the model were adopted from the work of Li et al. (2017) and H.-S. Song et al. (2017). The reactions of oxidative respiration and denitrification for the dissolved organic carbon (CH$_2$O) and the synthesis of biomass (C$_5$H$_7$O$_2$N) were considered as follows:

$\begin{align}
\text{CH}_2\text{O} + \text{O}_2 &= \text{CO}_2 + \text{H}_2\text{O} \\
\text{CH}_2\text{O} + 2\text{NO}_2^- &= 2\text{NO}_3^- + \text{CO}_2 + \text{H}_2\text{O} \\
\text{CH}_2\text{O} + 4/3\text{NO}_3^- + 4/3\text{H}^+ &= 2/3\text{N}_2 + \text{CO}_2 + 5/3\text{H}_2\text{O} \\
\text{CH}_2\text{O} + 1/5\text{NH}_4^+ &= 1/5\text{C}_5\text{H}_7\text{O}_2\text{N} + 3/5\text{H}_2\text{O} + 1/5\text{H}^+. 
\end{align}$

We used an enzyme-based cybernetic modeling approach to represent biogeochemical reaction regulated by microbial activities (H. Song et al., 2017; X. Song et al., 2018). Temperature-dependent reaction rate, $r_i$, is defined using the Arrhenius equation, that is,

$$r_i = r_{\text{base},i} \exp \left[ -\frac{E_a}{R} \left( \frac{1}{T} - \frac{1}{26 + 273.15} \right) \right], \quad i = 1, 2, 3,$$

where $r_i$ is the reaction rate under temperature $T$ (in Kelvin) with subscript $i$ being the generic reaction index, $E_a$ is the activation energy (0.65 eV in this study), $R$ is the gas constant (8.314 J·mol$^{-1}$·K$^{-1}$), and $r_{\text{base},i}$ is the base rate derived from laboratory batch experiments (under the room temperature of 26 °C). After incorporating microbial regulation term into the Monod kinetics, the base rate, $r_{\text{base},i}$, is defined by

$$r_{\text{base},i}^{\text{kin}} = e_i^{\text{rel}} k_i d_i/K_{a,i} + d_i K_{d,i} + a_i, \quad i = 1, 2, 3,$$

where the relative enzyme levels, $e_i^{\text{rel}} = r_{i}^{\text{kin}} / \sum_{j=1}^{3} r_{j}^{\text{kin}}$, regulate the rates of oxidative respiration and denitrification reactions (H. Song et al., 2017), $k_i$ (mol·L$^{-1}$·day$^{-1}$) is the maximum specific uptake rate of organic carbon (CH$_2$O), $a_i$ (mol/L) is the electron acceptor concentration, $d_i$ (mol/L) is the electron donor concentration, and $K_{a,i}$ (mol/L) and $K_{d,i}$ (mol/L) are the half-saturation constants for electron acceptors and electron donors, respectively. More details about the reaction network and rates can be found in X. Song et al. (2018).

**3.2. Uncertainty Sources**

We opt to include six important uncertainty sources: (1) climate scenarios, including time-varying river stage, inland well water level, and temperature; (2) the thickness of alluvial layer given a known shape; (3) heterogeneity or homogeneity formulations for the permeability field of Hanford Formation and the alluvial layer; (4) permeability fields in all of the formations no matter being homogeneous or heterogeneous; (5) reaction rates for $r_1$, $r_2$, and $r_3$ in equation (15); and (6) whether the heat transport process is modeled to enable temperature-dependent reaction rates. These uncertainty sources include all scenario, model, and parameter uncertainties: The climate scenario represents the scenario uncertainty, the heterogeneity or homogeneity of formations and with/without heat transport process represents different physical or
mathematical models thus can be categorized as model uncertainty, and the others are parameter uncertainty. Consequently, the BN in Figure 1 is simplified to include only the nodes that are relevant to this 2-D conceptual model as shown in Figure 3.

All six scenarios, one from each year of 2010–2015, were selected to start with similar river stage, inland groundwater table, and temperature at the beginning (Figure 4a) so that there is not significant uncertainty caused by the initial condition. Five different thicknesses were considered for the alluvial layer, and 100 different realizations were generated for the reaction rates and permeability fields combination using the Latin hypercube sampling for both heterogeneous and homogeneous within-formation permeability cases. The permeability field samples were generated using geostatistical simulations (Rubin, 2003) conditioned on point permeability measurements assuming an exponential variogram model. The reaction rates of \( r_1 \), \( r_2 \), and \( r_3 \) were assumed to follow uniform distributions ranging from 0 to twice of their mean values (28.26, 23.28, and 84.78 mol·L\(^{-1}\)·day\(^{-1}\), respectively), which were adopted from H.-S. Song et al. (2017) and used by X. Song et al. (2018). The total number of model simulations considering all possible combinations of model inputs and model configuration is \( 6 \times 5 \times 2 \times 2 \times 100 = 12,000 \), which represents six scenarios, five thicknesses of alluvium layer, heterogeneous/homogeneous formations, with/without heat transport process, and 100 reaction rates and permeability fields.

4. Results

4.1. Simulated River Tracer and Carbon-Consumption Rate

The concentration of river tracer and organic carbon consumption rate are the two primary outputs of our interest. The concentration of river tracer is the marker of the hydrologic exchange between the groundwater and river water, driven by the hydraulic gradient between the river and the underlying groundwater aquifer. The carbon-consumption rate is the most important indicator for the biogeochemical processes because it is the energy source for all biogeochemical reactions in the reaction network of the underlying system.

The hydrologic exchange is driven by the hydraulic gradient between the river and the groundwater aquifer. It is shown in Figure 4a that both the river stage and the inland groundwater table are constantly fluctuating, leading to dynamic hydraulic gradients that drive river water in and out of the groundwater aquifer. The
simulated river tracer concentrations in the aquifer are normalized by their concentration in the river \( (C_0) \), that is, \( C/C_0 \), and their snapshots at the end of the eighth, tenth, and twelfth weeks of simulation under scenario 1 are shown in Figure 4b for two distinct thicknesses of the alluvial layer. These later time points during the simulation window were selected to minimize the initial condition impacts and to capture...
longer residence time for biogeochemical reactions. The different geological formation boundaries are marked using black lines.

The river tracer and carbon-consumption rate vary in space and time as shown in Figures 4b and 4c, driven by the hydrologic forcing dynamics demonstrated in Figure 4a. The two examples with distinct alluvial layer thickness shared the same geostatistical realizations of permeability formation, that is, we first selected a realization for the larger spatial domain for each formation and masked the areas that fall outside of the formation for the other case. The river tracer plume first spread from the alluvial layer to the Hanford formation then started to retreat from the aquifer (Figure 4b), which led to similar spatial pattern for the carbon-consumption rate (Figure 4c) because the river provides the primary organic carbon source for all the biogeochemical reactions with oxygenated conditions on both ends of the domain. Due to the continuous consumption of organic carbon along the river water intrusion path, the carbon consumption rates kept reducing accordingly. Consequently, the zone that is biogeochemically active appeared to be smaller than the river tracer plume and the biogeochemical reaction rates in the Hanford formation were significantly lower than those in the alluvial layer. The low-permeability alluvial layer acts as a resistance layer to the river water intrusion that increases residence time for biogeochemical reactions. Thus, the alluvial thickness plays a significant role in controlling the spatial extent of river water intrusion. The snapshots associated with the case of a thick alluvium layer (Figure 4c) showed significantly more constrained extent of river water intrusion and biogeochemical activities compared to the case with thinner alluvial layer. The reaction front exhibited similarity to the geometry of the thick alluvial layer. There was negligible river tracer and carbon consumption in the Ringold formation because its low permeability prevents river water infiltration.

### 4.2. Sensitivity Analysis Results for Carbon-Consumption Rate

The sensitivity indices corresponding to four primary model input uncertainty sources at the process level, that is, climate scenario, groundwater flow, heat transport, and reactive transport, were calculated for the carbon-consumption rates at simulation times of 8, 10, and 12 weeks after the starting point using equation (7). The spatially distributed sensitivity indices are shown in Figure 5, including only the grid cells with a significant average carbon-consumption rate of $3 \times 10^{-8}$ mol·m$^{-3}$·s$^{-1}$ or more to avoid numerical artifacts.

Both the reactive transport and groundwater flow processes were found to contribute significant uncertainties to the carbon-consumption rate at all three simulation times in most of the biogeochemically active zones, with exceptions in small areas within the river stage fluctuation range where the saturation-desaturation dynamics controlled the biogeochemical reactions. Thus, the climate scenario contributed the most uncertainty in these areas, while it contributed the third largest portion of uncertainty to simulated biogeochemical reaction rates in general. The reactive transport process appeared to be more important in the area near the river-sediment interface where the carbon supply was less limited and flow is slower. In contrast, the flow process was more dominant in the areas away from river in the Hanford formation where the river water intrusion is faster. The heat transport process contributed little to the overall uncertainty, reflected by their small sensitivity indices across the model domain (<0.1). The thermal process was the least important uncertainty source for the carbon-consumption rate, indicating that the temperature-dependent reaction rates caused negligible difference in regulating the biogeochemical reactions in this carbon-limited system, compared to other processes. These sensitivity analysis results were supported by the simulated spatial and temporal dynamics of carbon consumption shown in Figure 4 that carbon consumption was driven by the river water intrusion, which supplied organic carbon for the biogeochemical reactions. Both the groundwater flow process and driving force scenarios interact to define velocity field dynamics as shown in Figures 1 and 3.

Within the groundwater flow process, its input uncertainty sources can be further attributed to two primary sources: geological formation structures and the permeability field within each formation, as described in section 2.2. Their corresponding sensitivity indices were computed via equation (13) and shown in Figure 5b for simulation times at 8, 10, and 12 weeks. The results illustrated that the geological formation structure contributed significantly more uncertainty to the flow than the within-formation permeability field in most of the modeling domain over time. The sensitivity analysis results on flow process components were consistent with the observation that the thickness of low-permeability alluvial layer plays an important role for river water intrusion and carbon source infiltration.
5. Conclusions

We developed a new sensitivity analysis framework that integrates the concept of BN- and variance-based global sensitivity analysis. We used a BN to capture the causal relations for a complex numerical model and represent its various uncertainty sources. The graphical representation of the model system clearly demonstrates all the causal relationships between the system behaviors (i.e., model outputs) and the various uncertain components, which greatly facilitates not only the understanding of a complex system but also communicating the scientific understanding to various stakeholders. Combined with a variance-based global sensitivity analysis that has been adapted to flexibly grouping variables through the BN, the new framework allows the identification of significant uncertainty contributors (a group of variables) in a hierarchical way beyond a restrictive three-layer structure. A new set of sensitivity indices on the model processes or their components was derived following the variance-based global sensitivity analysis method.

Figure 5. Spatial distribution of sensitivity analysis indices for (a) scenario ($S_S$), groundwater flow ($S_F$), heat transport ($S_H$), and reactive transport processes ($S_R$) at the simulation times of 8, 10, and 12 weeks. (b) Geological information ($S_G$) and permeability ($S_P$) subprocesses within the groundwater flow process at the simulation times of 8, 10, 12 weeks. The five thick black lines in the domain represent five distinct alluvium layer thicknesses. OC = organic carbon.
For test and evaluation purposes, the new sensitivity analysis framework was implemented for a real-world groundwater biogeochemical model, which simulated the hydrologic exchange flows and the associated reactive transport processes at the DOE Hanford Site's 300 Area. A BN was built for the biogeochemical model with six known input uncertainty sources grouped into four primary components at the process level. The sensitivity indices revealed that both the groundwater flow and reactive transport processes contributed significantly to the predictive uncertainty in carbon-consumption rate, followed by the climate scenario that defines the driving forces of the system. The heat transport process was much less important compared to the other three processes. Through further decomposition of the uncertainty contributed by groundwater flow process to the carbon-consumption rate, we found that the geological structural information was more important than the within-formation permeability field in controlling the flow process. The sensitivity indices at the process and subprocess levels all varied in space and time. Therefore, spatial and temporal sensitivity indices were essential to understand a dynamic and heterogeneous system.

Although the focus of this research is developing and testing a new sensitivity analyses framework suited for complex biogeochemical model, we also demonstrated through the use case that important knowledge and insights can be gained for improving general biogeochemical reactive transport model through sensitivity. It is worth mentioning that the BN is built on an acyclic graphic structure. For some natural processes that contain cyclic linkages, those nodes should be combined into one aggregate node in BN, such that the sensitivity contributed from this group could still be quantified although their individual contributions cannot be separated. The methodology in this research is also mathematically rigorous and generally applicable to a wide range of multiprocess earth system problems to reduce the predictive uncertainty by identifying the most important uncertainty contributors.

Acknowledgments
This research was supported by the U.S. Department of Energy (DOE), Office of Biological and Environmental Research (BER), as part of BER's Subsurface Biogeochemical Research Program (SBR). A portion of methodology improvement was supported by the Laboratory Directed Research and Development Program at Pacific Northwest National Laboratory, a multiprogram national laboratory operated by Battelle for DOE under contract DE-AC05-76RL01830. The third author was supported in part by the National Natural Science Foundation of China (NSFC) project 41807182. The third author was supported in part by DOE grants DE-SC0019438 and DE-SC0008272 and NSF EAR grant 1552329. This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science-supported User Facility under contract DE-AC02-05CH11231. The data used in the analyses can be downloaded online (https://sbrfa.velop.nrel.gov:443/data-sets/?CUID=79b70808-5F91-4d4a-a1b3-24c00586d4bc).

References


