Modeling the Evolution of Natural Organic Matter in the Environment with an Agent-based Stochastic Approach

Xiaorong Xiang Yingping Huang Gregory Madey
Department of Computer Science and Engineering
University of Notre Dame
Notre Dame, IN 46556
xxiang1@nd.edu, yhuang3@nd.edu gmadey@nd.edu
Steve Cabaniss
Department of Chemistry
University of New Mexico
Leilani Arthurs Patricia Maurice
Department of Civil Engineering and Geological Sciences
University of Notre Dame

Abstract

Natural organic matter (NOM) is ubiquitous in terrestrial and aquatic ecosystems, and it plays a crucial role in the evolution of soils, the transport of pollutants, and the global carbon cycle. NOM is a complex mixture of molecules and is thus heterogeneous in structure and composition. As NOM passes through an ecosystem, it is acted upon by a variety of processes, such as microbial degradation, adsorption to mineral surfaces, and photochemical reactions that can change its properties and reactivity. The evolution of NOM in space and time thus is an important research area in biology, geochemistry, ecology, soil science, and water resources. Due to its complex structural and chemical heterogeneity, new simulation approaches are needed to help better understand the evolution of NOM properties and reactivity as it passes through an ecosystem. We present a new stochastic model, which explicitly treats NOM as a large number of discrete heterogeneous molecules ("agents") with different probabilities of transformations or reactions. The NOM, the microorganisms, and their environment are taken together as a complex system, with the NOM interactions within this system simulated using an agent-based stochastic modeling approach. The initial users of the NOM simulations include a geographically separated group of NSF-sponsored scientists and engineers from different research disciplines, in-

cluding both academics and U.S. government scientists. A Web-based interface serves as a prototype NOM "collaboratory" designed to promote collaboration among the various researchers and to allow them to share their data, model results, and suggested approaches or improvements across distributed sites. This Web-based interface has been designed to allow researchers to access the simulation model remotely from a standard Web browser. The Web-based interface thus allows researchers at distant locations to provide parameters for their simulations, to start and stop simulations, and to plot and view results, all remotely.

1. Introduction

Natural organic matter (NOM) is a polydisperse mixture of molecules with different structures, compositions, functional group distributions, molecular weights, and reactivities, that forms primarily as the breakdown product of animal and plant debris, and that can also be derived from algal sources. NOM is ubiquitous in terrestrial and aquatic ecosystems, and has been widely reported in marine environments. Its structure, chemical composition, and reactivity vary both spatially and temporally. NOM plays a crucial role in ecological and bio-geochemical processes such as the evolution of soils, the transport of pollutants, and the global bio-geochemical cycling of elements [6]. NOM is a primary food source to microorganisms, and it can act as a natural 'sun block,' attenuating potentially damaging UV radiation [26] in lakes and streams. While passing through an ecosystem, NOM may be acted upon and potentially altered by a wide array of processes, such as microbial biodegradation, adsorption to mineral surfaces, redox reactions, photochemical processes, and aggregation or coagulation. The evolution of NOM over space and time from precursor molecules to eventual mineralization (primarily as CO_2) is an important research area in a wide range of disciplines, including biology, geochemistry, ecology, hydrology, soil science, and water resources. Given the widespread abundance and importance of NOM to many hydro- bio- geochemical processes, predictive modeling of its evolution in structure, composition, and reactivity are fundamental to many areas of environmental research.

Because of the complex nature of NOM, and the multitude of possible reactions it may undergo in natural environments, we have only a limited knowledge of the detailed mechanisms by which it forms from precursor molecules, or how its structure, composition, and reactivity evolve over space and time. Perhaps most notably, the fact that NOM is a polydisperse mixture of molecules which themselves have complex structures means that it defies characterization by established analytical methods. Previous models of NOM formation and evolution have been important for predicting certain types of reactions or interactions, but they have not beeen able to describe both the quantitative aspects of organic carbon transfer and the semi-quantitative or qualitative aspects of NOM structure and functional heterogeneity. Predicting NOM interactions and their consequences to other environmental processes requires a clear understanding both of how a single NOM component behaves

and how the entire NOM mixture at a given site evolves over space and time.

In the previous carbon cycling models for NOM study, NOM is treated as a single "organic carbon" entity, and the "average" values of properties is used to represent the complex NOM mixture [28] [13] [5] [12]. The use of average properties of various organic carbon pools when modeling NOM is too simplistic to represent the heterogeneous structure of NOM and its complex behavior in the environment. Also, it can result in discrepancies when the NOM model results are compared with the results from the laboratory studies. The other models that employ specific chemical structures and reactions (e.g., ab initio quantum mechanics and molecular mechanics approaches) are too computationally intensive to be useful for large-scale environmental simulations [25].

In this paper, we present a novel middle approach, an agent-based stochastic model as shown in Figure 1, that explicitly represent individual molecules as discrete objects with partial structural and functional properties. The temporal evolution of NOM is modeled using a Monte Carlo method in which specific probabilities are assigned to particular transformations. The reactivity of the resulting NOM over time is predicted based on the distributions of molecular properties.

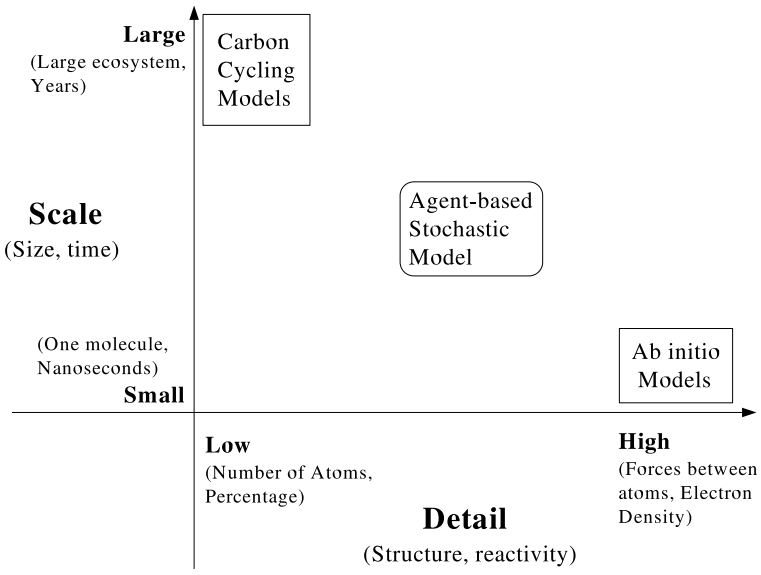

Figure 1. Agent-based stochastic model is the middle approach

The Monte Carlo method is relatively easy to program and can be quite flexible, since it does not depend upon the availability of analytical solutions. Also in the stochastic

approach, complex behaviors of NOM are derived from simple rules and probabilities. It is often simpler to program than deterministic approaches when simulating the dynamics of a complex system.

Agent-based modeling (ABM), also known as individual-based modeling (IBM), is a method used to track the actions of multiple agents that can be defined as objects with some type of autonomous behavior [17] [1] [21] [10]. By using the ABM approach, the higher level behaviors of a system, called "emergent behaviors" of the system, can be discovered without being explicitly coded into the simulation. Predicting the phenomena at higher levels is based on the actions of individual agents in a "bottom-up" approach. The technique of building and using ABMs is a useful tool for understanding complex systems and is increasingly applied to ecological, economic, environmental, and sociological research. An alternative approach to modeling a system is equation-based modeling (EBM), also called "state variable" approach [10], involves solving a set of differential equations. EBMs can describe already known global properties of a system in a "top-down" approach, but often can not explain the origin of those properties, track the behavior of individual components, or represent extreme heterogenity. ABMs and EBMs are significantly different with respect to which characteristics they focus on. ABMs focus on the characteristics of each individual and track them through time. Equation-based models, on the other hand, focus on the characteristics of the population, which are averaged, and simulate changes in the averaged population characteristics [20]. Grimm [10] argues that the two approaches are complementary and both provide insight and can potentially both add to scientific theory.

Van Dyke Parunak [20] working in the domain of supply networks, presented an agent-based model and an equation-based model for modeling a supply network, and concluded that an agent-based model is more suitable for modeling complex systems that are composed of interacting components and exhibit a wide range of dynamic behavior.

The NOM, micro-organisms, and their environment are taken together as a complex system with wide range of dynamic behaviors. With the agent-based stochastic model, individual components (molecules), modeled as individual agents, are given a set of simple rules on how to move through soil pores and how to interact with each other. The changes in individual components or a group of agents, which start with the same properties over time, can be tracked. The global properties (including distributions of physical, chemical, and biological properties) of NOM evolution over time can be predicted by simulating the physical and chemical reactions between individual agents with temporal and spatial properties.

The advent of the Internet and advanced digital computer technology has produced a new generation of Web-based applications [9]. Web-based scientific applications use the Web as a new platform to do research by combining recent information technologies and computational approaches. Unlike other simulators used in the study of NOM thus far, the NOM simulators are deployed as Web-based scientific applications that are mainly focused on end users and collaboration. They offer scientists the opportunity to share their data and information with others in the research community by providing a suite of collaboration tools via the Internet. Scientists can access these simulators through the Web interface,

configure a particular simulation by providing a set of parameters, and query the data stored in the online databases. This new infrastructure, has been used to solve problems such as distributed physical or astronomy data analysis [19], and remote access of the information sources and simulations [8][15]. By taking advantage of the database and data mining technologies, large data sets can be stored and the changes of the system can be traced.

In this paper, the conceptual model of the evolution of NOM is presented in section 2. Five computer implementations or simulations to date based on the conceptual model are described in section 3. The Web-based interface is illustrated in section 4. We discuss preliminary verification and validation of the model and simulations in section 5. The concluding remarks are drawn in section 6.

2. The Conceptual Model

The conceptual model described in this section defines our theoretical world view and ontology that guides the actual design and implementation of computer simulations described later in this paper. Based on the observational and experimental studies on the behavior of NOM in the environment, we designed a conceptual agent-based stochastic model. In this model, NOM, micro-organisms, and their environment form a complex system. The evolution of NOM over discrete time and space from precursor molecules (such as cellulose, lignin, and protein) to eventual mineralization involves various molecular transformations. These transformations involve chemical reactions, adsorption, aggregation and physical transport in soil, ground, or surface waters. NOM is presented as a large number of discrete molecules with varying chemical and physical properties. Individual molecules, modeled as individual agents, are given a set of simple rules on how to move through soil pores and how to interact with each other.

By modeling and simulating the behaviors of individual molecules in the system, we expect that the distributions of physical, chemical, and biological properties of NOM can be predicted. Additionally, such simulations can provide scientists in biology, geochemistry, and ecology valuable insight into the processes underlying complex environmental phenomenon.

2.1. Agents

In the NOM complex system, the agents are NOM molecules. These NOM molecules are hypothesized to be derived from macromolecules such as proteins, polynucleotides, cellulose, lignin, or small organic molecules such as phospholipids, sugars, and amino acids. To avoid dealing with each possible macromolecule, we work with "representative" molecules, which have typical structures and elemental compositions. These representatives can be used to obtain reasonable similarity to the actual macromolecular precursors.

It is impractical to define precise molecular representations that involve descriptions of atomic location, electron density maps, and force field constants given our goal of modeling

thousands or even up to one million individual molecules. We therefore need an intermediate level molecular representation as shown in Figure 1, which is more specific than simply "percent carbon" but less detailed than a precise molecular connectivity map. Also, the representations of structures should be detailed enough to illustrate the heterogeneity of the NOM.

The data used for constructing representations of NOM molecules includes the following components:

- Elemental formula, i.e. the number of **C, H, O, N, S, P** atoms in the molecule. The molecular weight **MW** can be easily calculated from the elemental formula.
- Functional group count, e.g. **carboxylic acid, alcohol, ester groups**. There are a total of 19 possible functional groups in our model for each molecule structure.
- A record of the molecular "origin," i.e. the initial molecule, its starting position, and its time of entry into the system. This allows for the calculation of separate "turnover times" and apparent ages for individual molecules or fractions. The location of each molecule in the simulation is represented by the x and y coordinates of a 2-dimensional lattice.

2.2. Behaviors

In the environment, NOM can move with percolating water down a soil column and into the groundwater, and interact with other molecules and their environment. These interacting behaviors can be separated as physical reactions or chemical reactions. The key distinction between these two is whether the molecular attributes change or whether the position/state of the molecule changes. The physical reactions modeled are sorption to and desorption from mineral surfaces, as molecules are transported by water through soil pores. Sorption is not a transformation of molecular structure; however, it will affect the probability of other reactions by changing the environment of a given molecule.

In this conceptual model, the individuals are associated with a location in geometrical space and can move around their environment. The geometrical space is described as a discrete 2D grid space represented by integer values. Molecules enter the system at the top of the grid space, move with the water flow, and leave the system at the bottom of the grid.

The physical reactions are modeled based on the results of laboratory experiments. As described by Zhou et al.[30], the lower molecular weight molecules adsorb and desorb quickly and are gradually replaced at surface sites by larger molecules that adsorb and desorb slowly. This correspondence between empirical data and model behavior is used to help validate the simulation. This is discussed in more detail in section 5.

Chemical reactions result in structural changes in the molecule, such as the addition of groups to a NOM molecule. New molecules can be generated from the predecessor molecules and those predecessor molecules may leave the system. Twelve types of chemical reactions, including first-order and second-order chemical reactions, are modeled as described in Table 1. Others will be added as needed.

Table 1. Chemical Reactions in the Conceptual Model

Reaction Name	Reaction Type
Ester condensation	Second order
Ester hydrolysis	First order with split
Amide hydrolysis	First order with split
Microbial uptake	First order with molecule disappear
Dehydration	First order with split
Strong C=C oxidation	First order with split (50%) of the time
Mild C=C oxidation	First order without split
Alcohol(C-O-H) oxidation	First order without split
Aldehyde C=O oxidation	First order without split
Decarboxylation	First order without split
Hydration	First order without split
Aldol condensation	Second order

These twelve chemical reactions are separated into four categories:

- First order reactions with a split are defined as follows: The predecessor molecule **A** is split into two successor molecules **B** and **C**, molecule **B** occupies the position of molecule **A**, and one of the empty cells nearest molecule **B** is filled with molecule **C**.
- First order reactions without a split are defined as follows: The transformation only changes the structure of the predecessor molecule **A**.
- First order reactions with the disappearance of a molecule are defined as follows: The predecessor molecule **A** disappears from the system. This reaction occurs when the molecule is small enough that it can be enveloped by a micro-organism like fungi or bacteria.
- Second order reactions are defined as follows: Two molecules **A** and **B** are combined to form a new molecule **C**; the new molecule **C** replaces molecule **A** while the other predecessor molecule **B** disappears from the system.

  During each time step, molecule **A** is chosen. If the generated random number indicates that a second order reaction should occur, a second molecule **B**, which is nearest the location of **A**, is found in the grid. For example, for an ester condensation reaction, molecule **B** should have non-zero alcohol groups. These two molecules combine to form Molecule **C**. The probabilities are calculated for molecule **C** and the structure for molecule **C** is determined. Molecule **C** replaces the position of **A** and molecule **B** leaves the system.

2.3. Reaction probabilities

The common rate law governing macroscopic chemical kinetics is that for first order reactions in which the rate or velocity of a reaction **R** (units of concentration per time) depends only on the concentration of the reacting molecule **C** (units of concentration) and a rate constant **k** (units of reciprocal time).

$$R = kC \tag{1}$$

In the second-order reactions, the rate depends on the concentration of two reacting molecules C_1 and C_2 as shown in equation 2. It can be reduced to an apparent first order equation 3.

$$R = kC_1C_2 \tag{2}$$

$$R = k^{'}C_1 \text{ where } k^{'} = kC_2 \tag{3}$$

The probability for each reaction type is expressed in terms of intrinsic and extrinsic factors. Intrinsic factors are derived from the molecular structure including the number of functional groups and any other structural factors. Extrinsic factors arising from the environment include concentrations of inorganic chemical species, light intensity, availability of surfaces, presence of micro-organisms, presence and concentration of extracellular enzymes, and the presence and reactivity of other NOM molecules. The intrinsic and extrinsic factors are combined in probabilistic functions that are defined by Cabaniss [6] for a particular molecule under a particular set of circumstances. Each set of **N** reactions for a particular molecule can occur with some probability **P**.

The probability that a molecule will react via reaction i over the time interval Δt is defined in equation 4 where $k^{'}$ is the first order or apparent first order rate constant.

$$P_{i,\Delta t} = \Delta t k_i^{'} \tag{4}$$

The environmental parameters that affect the apparent rate constants include acid and base concentrations (i.e. pH), temperature T, dissolved oxygen gas (O_2), light density (I), microbial density (MD), fungal density (FD), and water ($Water$).

The probability of each reaction incorporates the effects of oxidizing reactions, photochemical reactions, and biochemical reactions. For example, the probability of the **Ester hydrolysis** reaction for a given molecule depends on the number of Ester functional groups $\#Ester$ in the molecule and can be calculated using equation 5. In this equation, A, b, c are constants, E_a is activation energy, and R is ideal gas constant.

$$P_{esterhydrolysis} = \Delta t(Water)(\#Ester)Ae^{\frac{-Ea}{RT}}(1 + b[H^+] + c[OH^-]) \tag{5}$$

$$\text{where } A = 6 \times 10^5 hr^{-1}, E_a = 60 kJ mol^{-1}, b = 10^4, c = 3 \times 10^8$$

2.4. Molecular properties

The reactivity of the resulting NOM over time can be predicted based on the distributions of molecular properties. Some molecular properties of NOM can be calculated and predicted. These properties are quantities which can be calculated from the elemental composition and functional group data. They represent a measurable quantity which can be used as a predictor for an environmental function. They are also useful and interesting both as part of scientific inquiry and for the calibration and verification of our conceptual model and simulation. Some properties are easy to calculate, such as molecular weight, molecular charge, and charge density. On the other hand, others involve non-trivial computations. Also, some properties are exactly calculated (e.g., molecular weight) while others must be estimated by empirical relationships which will introduce additional uncertainty (e.g., pKa).

2.5. Simulation process

The conceptual model is a stochastic synthesis model of NOM evolution. It serves as a design framework for the computer simulations described later in this paper. In a stochastic process, the state of the system is represented by a set of values with a certain probability distribution, such that the evolution of the system is dependent on a series of probabilistic discrete events.

Individual molecules are represented as agents with their own complex structures. These molecules move along with the water flow and react with each other in a 2D discrete grid, i.e. a rectangular lattice composed of multiple cells. Each molecule can occupy at most one cell and each cell can host one or more molecules. During each time step, each molecule may move to another location, experience adsorption to or desorption from a particular site, and chemically react with other molecules. The basic algorithm for building simulations from the conceptual model is illustrated in Figure 2. It includes the following major steps (not all may be used in any particular simulation):

1. Initial molecule creation: Molecules are created by obtaining values for all quantities in the molecule's structure. The quantities are obtained from users' inputs and calculated from these inputs.

2. Modeling the movement and sorption (for porous media): As molecules move along with the flow, they can be adsorbed to surfaces or desorbed from them based an random probabilities.

3. Modeling the reaction: Molecules are allowed to react for a predetermined length of time under fixed external conditions and molecule structures are recorded into the database at defined time increments.

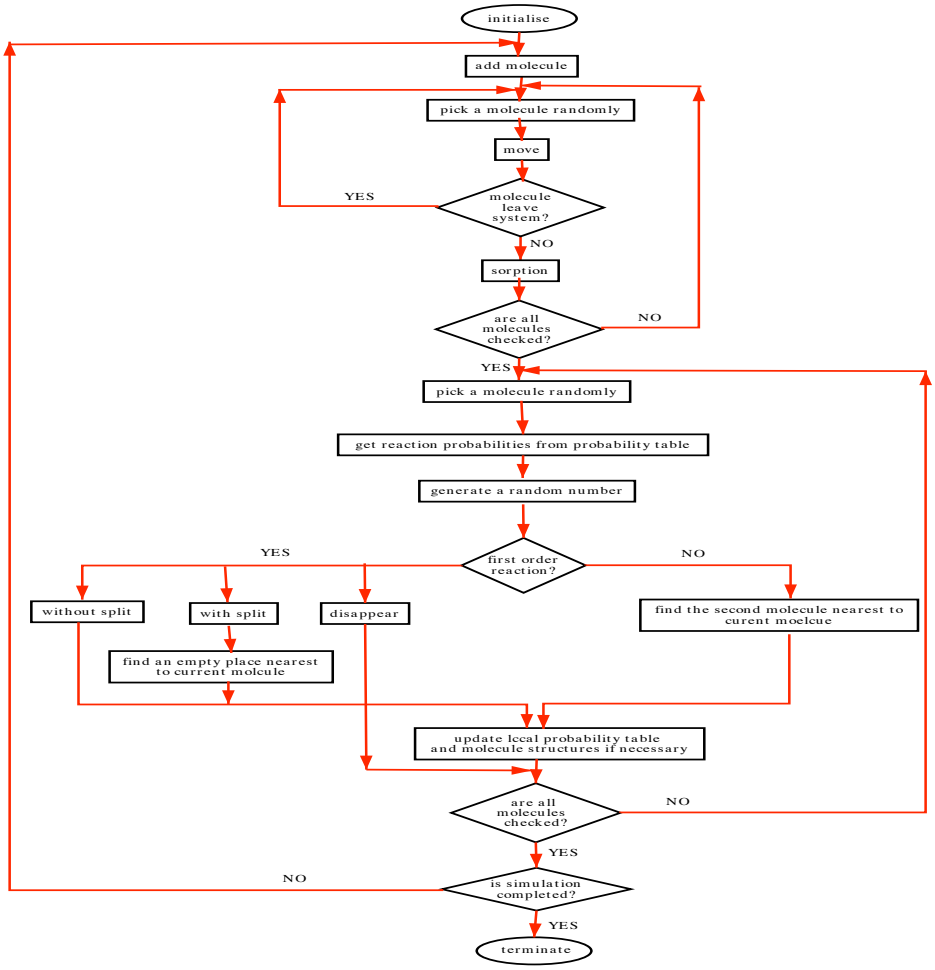

Figure 2. A flow chart for building simulations from the conceptual model. Actual implementations of NOM computer simulations may include a subset of these steps since they focus on a one aspect of NOM behavior, e.g., adsorption, or reactivity.

4. Property calculation: In order to better understand and monitor the global properties of NOM over time, large amounts of information for the system are stored in the database for calculation and analysis.

The simulation time, defined by the user, is divided into a very large number of equal, discrete, and independent steps called "time steps." In each time step, each molecule in the system is chosen in random order. A random number is generated to see which reaction will occur. To determine which reaction will take place in a given time step, we calculate the probability p_i of each reaction according to the environmental variables and the structure of the chosen molecule. This random number is used to compare with the precalculated reaction probabilities associated with the chosen molecule. The molecule structure and properties associated with the molecule are updated after each physical or chemical reaction occurs.

The sum of all the reaction probabilities is controlled to be less than 1 percent base on the assumption that each molecule has at most one reaction per time step. This is controlled by choosing sufficently short time steps, ΔT, e.g., see equation 5. However, if the sum is too small, it will take longer time to simulate the process. Thus, in our model, the sum of all the reaction probabilities is set as:

$$\sum_{i=1}^{m} p_i < 0.01 \text{ where } m = \text{ number of reactions }, p_i = \text{ probability of reaction } i$$

The probability that two reactions occur in a time step for one molecule is less than $(0.01)(0.01) = 0.0001$. This number is small enough that this situation can be ignored in the model and higher numbers of reactions per molecule per time step are even much less likely.

The interval, [0, 1], is partitioned into 13 subintervals as shown in Figure 3. The length of the first interval is equal to the probability of the first reaction type; the length of the second interval is equal to the probability of the second reaction type, and the length of the ith interval is equal to p_i. The length of the last interval is the probability in which no reaction will occur.

$$p_{m+1} = 1 - \sum_{i=1}^{m} p_i > 0.99$$

The random number from the interval [0, 1] resides in one of these intervals, and it will decide which chemical reaction will occur if there is one.

At each time step, for each molecule, a uniform random number is generated. It is this number that determines whether a reaction will occur and if one does occur, it determines the reaction type. After a reaction takes place, the attributes for the current molecule are updated and reaction probabilities are recalculated. The molecule structure is changed to affect the outcome of the reaction and a new probability table entry is added for newly formed molecules, if there are any.

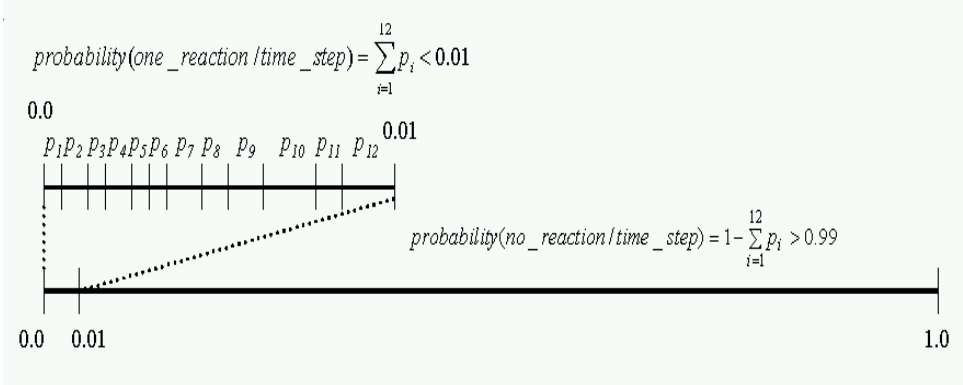

Figure 3. Probability calculation for each molecule at each time step

3. Implementations/Simulations

There are five implementations derived from the conceptual model as shown in Figure 4.

AlphaStep is a reference implementation that is coded in Delphi 6 and runs under Windows XP. It is a demonstration of the comprehensive conceptual model that doesn't have web and collaboration features. AlphaStep simulates a variety of chemical and biological transformations, but does not simulate any type of transport and does not represent the spatial properties of NOM. Hence, it represents a batch or closed system without porous media. AlphaStep is intended as a stand-alone application to allow ecologists, geochemists and environmental scientists to explore possible routes of NOM transformation. It is available for download from http://www.nd.edu/ nom/software.

The other four implementations are coded using the Java programming language (Sun JDK 1.4.2) and Swarm [27] [18] and Repast[22] software. Swarm is a software package for simulating complex systems that was developed at the Santa Fe Institute. It is a set of libraries that facilitates the implementation of agent-based models. RePast is a Swarm-like agent-based simulation toolkit written in the Java language. It was developed by the University of Chicago and Argonne National Laboratory. Swarm and Repast provide versatile random number generators and distributions, both essential to stochastic computer simulations. Also, these toolkits provide high level visualization capabilities. A control panel allows a running model to be stopped, restarted, or executed step by step. They also allow modelers to visualize the current state of the running model and probe every object in the model. An animation window shows the location of individual agents in a space. Line graphs and histograms are used to illustrate changes in collections of model objects

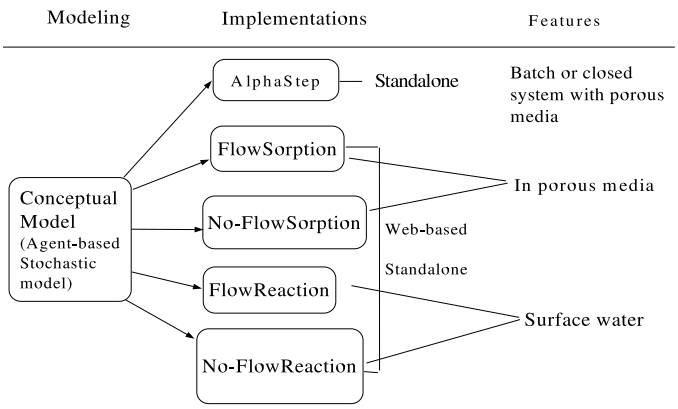

Figure 4. Five implementations derived from the conceptual model

that occur during the simulation. Additionally, simulation data, such as population size over time, can be reported periodically as the model executes. Figure 5 is a screen shot of a **FlowSorption** simulation used by the authors for verification and validation of the simulation.

The **FlowSorption** and **No-flowSorption** implementations respectively model laboratory batch and column adsorption experiments. In these two implementations, we focus on NOM molecular weight as a controlling feature on adsorption [7]; chemical reactions among molecules that are defined in the comprehensive conceptual model are not simulated to date. Groups of NOM molecules are represented as agents (different agents are defined by different molecular-weight intervals). Molecules can move, adsorb, and desorb in porous media according to a simple set of behavioral rules. The No-flowSorption simulates the batch adsorption experiment with all molecules added to the system at the beginning and diffusing throughout the medium. There are no molecules entering into or leaving the system during the subsequent time steps. On the other hand, the FlowSorption is characterized by the incremental input of molecules over time and transport down the column, with adsorption or desorption to mineral surfaces. Molecules can flow into and out of the modeled column system, resulting in a continuously changing population of molecules in the systems. Molecules undergo advection and dispersion.

The **FlowReaction** and **No-flowReaction** are implementations that simulate all the features described in the conceptual model. NOM molecules are represented as agents with properties defined by users. Molecules can move and interact with each other. Molecules

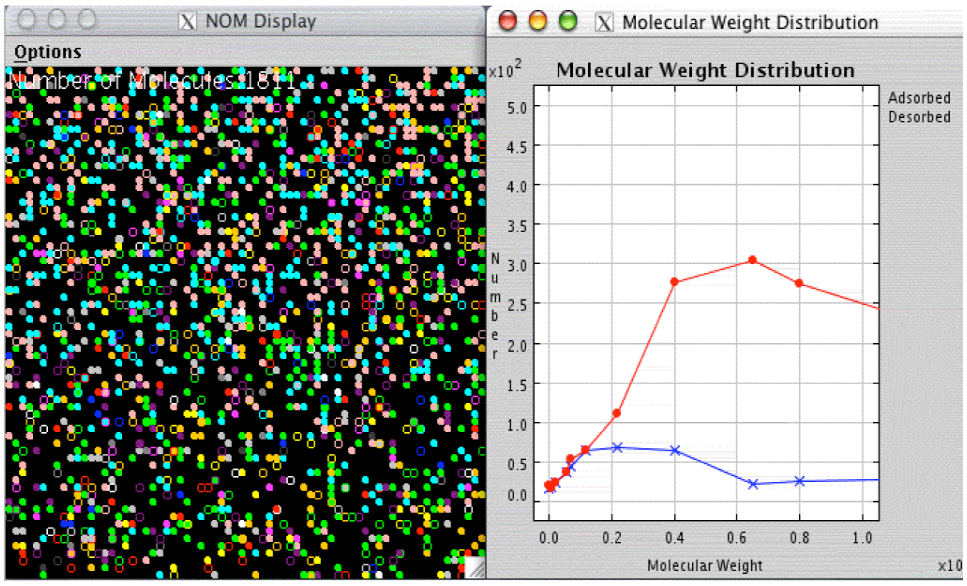

Figure 5. Example of a snapshot from a FlowSorption simulation. Left side displays molecules moving through the column: adsorbed (closed circles) or in solution (open circles). In a colored version, colors represent different MW intervals. Right side displays the corresponding MW distribution of adsorbed (higher peak) versus in-solution molecules. Adsorption is preferential for intermediate-to high MW components. In this example, the model input was molecular weight distribution of the NOM with adsorption controlled by molecular weight. Specific chemical reactions were not explicitly considered.

can be all uniformly distributed on the 2D grid at the beginning followed by diffusion or they can be added at a constant rate into the system to simulate their mobility with water flow including advection and dispersion.

There are two running modes for these four implementations: stand-alone mode coupled with a graphical user interface (GUI) and Web-based mode with Web-based interface. The GUI version is mainly used by the software developers for debugging, testing, and validation purposes, although it can be download and installed by users on their local computer. These four implementations are also deployed on the Web so that users can access them through a browser, such as Internet Explorer and Netscape. Section 4 focuses on the Web-based interface implementation.

4. Web based interface

Unlike other simulators used in the study of NOM thus far, our NOM simulators are built as Web-based applications. Huang [14] gives the detail of the infrastructure and database. Web-based applications are mainly focused on end users and collaboration. Compared with the traditional stand-alone application, there are several advantages of using the Web as a new platform for scientific study. Scientists do not need to download, build, and install the application packages or related software on their own computers which can sometimes be a tedious task, especially for applications written in Fortran, Pascal, C or C++ that need to be compiled to native code. Scientists can share the expensive computational resources or instruments, such as large-scale databases, which are not readily affordable for small groups. Web-based applications offer scientists the opportunity to share their data and information with others in the research community by providing a suite of collaboration tools. Deployment of the simulators on the Web promotes their use as a "collaboratory" or a collaborative laboratory for geographically separated scientists and engineers. In order to support collaborative work, a NOM collaboratory is built based on Sun Java 2 Enterprise Edition (J2EE) and relational database technologies, in particular the Oracle database. Our NOM collaboratory includes Web-based simulators, data analysis packages, simulation configurations, and communication tools such as discussion board and chat room. Xiang [29] provides details on the online collaboration features that can support environment researchers

Scientists can access these Web-based simulators through a Web browser. They choose a particular simulator and provide the input parameters that are then stored in the remote database. After they submit their configuration, simulations are invoked at a remote computer. When simulations are finished, users are notified by Email. They then can access the data results from the referred site. These data results not only include the raw data but also the graphic results that are generated by the data analysis packages using data mining technologies as shown in Figure 6. The built-in functionalities of our NOM collaboratory allow scientists to share all their simulation results, data, and information with others. They are also able to access data from previous simulations.

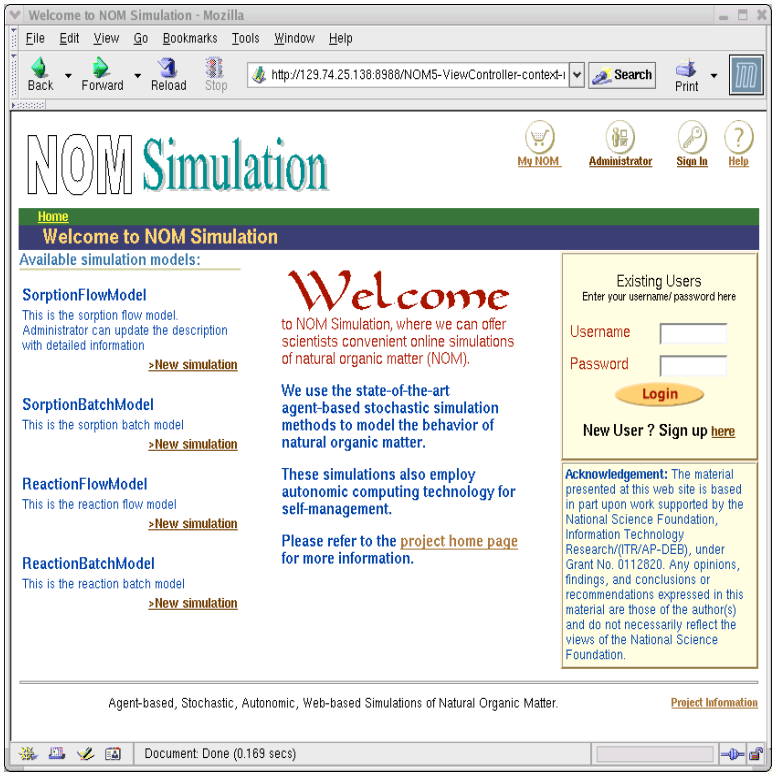

Figure 6. Screen shot of the Web based interface

5. Verification and Validation (V&V)

Verification and validation (V&V) are processes used to increase confidence in simulations. Verification is about getting the ”simulation right” while validation is about getting the ”right simulation”. Although neither process guarantees absolute confidence, we used numerous V&V techniques on the NOM simulation. We describe these techniques using an adapted version of Sargent's V&V process shown in Figure 7 [24].

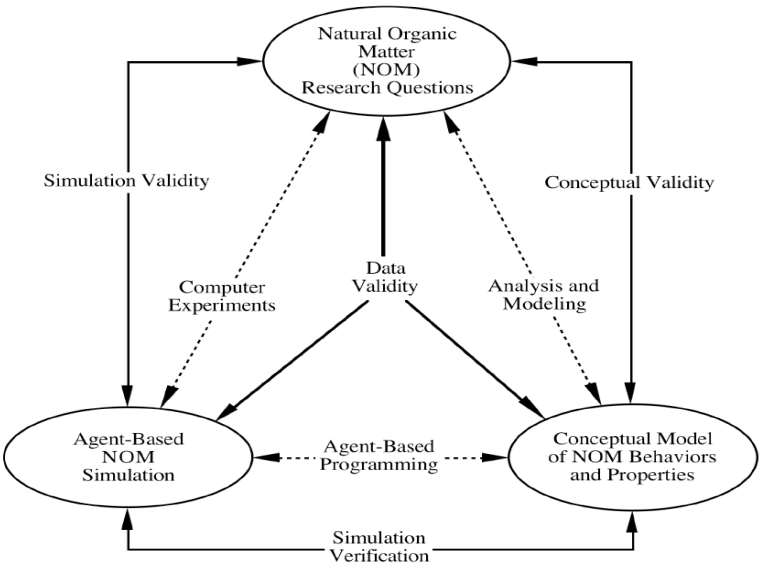

Figure 7. Verification and Validation Process

5.1. Conceptual Validity

The simulation process starts with an identification of research questions of interest. Through analysis and modeling, a conceptual NOM model is developed that includes the important features relevant to the research questions. The conceptual model is based on theory and domain knowledge from environmental chemistry, soil science, and geomicrobiology. This theoretical knowledge which guided the model development includes the following, as examples: 1) the heterogeneity of NOM molecules, 2) the important NOM interactions with mineral surfaces such as adsorption, hemi-micelle formations, acid or complexing dissolution, and reductive dissolution, 3) NOM interaction with pollutants, 4)

relationships between NOM adsorption to mineral surfaces and the molecular weight of the NOM molecules, and 5) probabilistic reaction kinetics based on elemental composition and the nature of functional groups in the molecules (e.g., 5). The incorporation of such theory and domain knowledge provides us initial face validity, ie., the logic of the conceptual model appears to domain experts to include appropriate mechanisms and properties of the research problem. Six scientists on the project - two biologists, a chemist, a geomicrobiologist, and two soil scientists - evaluated the conceptual model for face validity.

5.2. Simulation Verification

Once the conceptual model achieved its initial validation, coding of the agent-based simulation took place. In this step, verification methods such as code walk through, trace analysis, input-output testing, pattern tests, and boundary testing were used to verify the correctness of the simulation [3] [23]. Since the simulations are stochastic, the random number generators were confirmed to be robust and subjected to tests for uniformity and independence [4] [16]. When more than one implementation of the conceptual model is available, a cross-simulation comparison test can be used to verify both simulations. We developed five independent simulations with different features, using a formal model specification. One implementation, AlphaStep, is a stand alone simulation written in Pascal that runs on a PC; the other four implementations, are written in Java, uses agent simulation toolkits, and runs as a web-based application on Linux servers. This cross-simulation comparison, sometimes called docking or model-to-model comparison, was used to confirm corresponding behaviors between the two simulations, increasing verification confidence in both implementations.

5.3. Simulation Validity

To date, the validation of the simulation has included comparisons of simulation behavior with mathematical models and experimental laboratory results. For example, the simulation has been used to study the relationship between the adsorption of NOM molecules on mineral surfaces and the molecular weight of the molecules. NOM adsorption is preferential to intermediate to high molecular weight components, shown in laboratory data and confirmed in the field. The simulation has yielded a similar distribution as reported by Arthurs et al [2]. Additional such comparisons between theoretical, empirical and simulation predictions are planned.

Visualization of the behavior of the simulation is another useful technique used for simulation validation [11]. A snapshot of an animated visualization of the flow of molecules through a soil column depicting the adsorption and desorption of the molecules to mineral surfaces in the simulation is shown in Figure 5. In addition to the color coding of molecules by molecular weight, the adsorbed or desorbed states are depicted by solid and hallow circles respectively. A corresponding animated graph of the molecular weight distribution shows how the molecular weight distribution shifts with time: initially favoring

lower weight molecules in the early stages of the simulation and gradually shifting to larger molecular weights as the simulated time passes. These same behaviors were observed in laboratory experiments, increasing the confidence in the simulation.

6. Conclusion

In this paper, a new modeling approach, agent-based stochastic modeling, is presented for studying the behavior and evolution of NOM with temporal and spatial properties. Five simulators are derived and implemented based on the comprehensive conceptual model. Some implementations are partially validated by comparing with the laboratory data. More validation and verification is planned. Also four implementations are deployed on the Web to facilitate the data information sharing and collaboration among the geographically separated scientists. Eventually, we expect that the model can provide scientists with a test bed and predict the evolution of NOM over time.

Acknowledgements This research was supported in part by a NSF ITR Grant No. 0112820 and by the Center for Environmental Science & Technology at the University of Notre Dame.

References

[1] J. J. Anderson. An agent-based event driven foraging model. *Natural Resource Modeling*, 15(1), 2002.

[2] L. Arthurs, P. A. Maurice, X. Xiang, R. Kennedy, and G. R. Madey. Agent-based stochastic simulation of natural organic matter adsorption and mobility in soils. In *Eleventh International Symposium on Water-Rock Interaction*, June 2004.

[3] O. Balci. *Handbook of Simulation: Priciples, Methodology, Advances, Applications, and Practice*, chapter Verification, Validation and Testing. John Wiley & Sons, New York, 1998.

[4] R. Bowman. Evaluating pseudo-random number generators. *Computers & Graphics*, 19(2):315–345, 1995.

[5] G. Brown, S. Cabaniss, P. MacCarthy, and J. Leenheer. Cu(II) binding by a ph fractionated fulvic acid. *Anal. Chim. Acta 402*, 1999.

[6] S. E. Cabaniss. Modeling and stochastic simulation of nom reactions, working paper. http://www.nd.edu/ nom/Papers/WorkingPapers.pdf, July 2002.

[7] S. E. Cabaniss, Q. Zhou, P. Maurice, Y.-P. Chin, and G. Aiken. A log-normal distribution model for the molecular weight of aquatic fulvic acids. *Environmental Science and Technology 34*, pages 1103–1109, 2000.

[8] A scientific web-based application for global tropical cyclone monitoring. http://www.cio.noaa.gov/hpcc/projects/200128.html.

[9] G. Fox. E-science meets computational science and information technology. *Computing in Science & Engineering*, 4(4):84–85, July/August 2002.

[10] V. Grimm. Ten years of individual-based modelling in ecology: what have we learned and what could we learn in the future? *Ecological Modeling*, 115:129–148, 1999.

[11] V. Grimm. Visual debugging: A way of analyzing, understanding and communication bottom-up simulation models in ecology. *Natural Resource Modeling*, 15(1):23–38, Spring 2002.

[12] B. Gu, J. Schmitt, Z. Chen, L. Liang, and J. McCarthy. Adorption and desorption of different organic matter fractions on iron oxide: Mechanisms and models. *Environ. Sci. Technol. 28*, pages 38–46, 1995.

[13] S. Hansen, H. Jensen, and N. Nielsen. Daisy - a soil plant atmosphere system model. *NPO research from the National Agency of Environmental Protection*, 1990.

[14] Y. Huang, X. Xiang, G. Madey, and S. E. Cabaniss. Agent-based scientific simulation using java/swarm, j2ee and rdbms technologies. *Computing in Science & Engineering*, To appear.

[15] R. M. Jakobovits, J. F. Brinkley, C. Rosse, and E. Weinberger. Enabling clinicians, researchers, and educators to build custom web-based biomedical information systems. In *AMIA Annual Fall Symposium*, pages 279–283, 2001.

[16] P. L'Ecuyer. *Handbook of Simulation: Priciples, Methodology, Advances, Applications, and Practice*, chapter Random Number Generation. John Wiley & Sons, New York, 1998.

[17] J. G. Louis. Agent-based modeling in ethnobiology: A brief introduction from outside. http://www.tiem.utk.edu/ gross, April 2002. Talk to NSF Workshop on Priorities in Ethnobiology.

[18] N. Minar, R. Burkhart, C. Langton, and M. Askenazi. The swarm simulation system: A toolkit for building multi-agent simulations. Technical report, Santa Fe Institute Working Paper 96-06-042, 1996.

[19] US national virtual observatory. http://www.us-vo.org/software.html.

[20] H. V. D. Parunak, R. Savit, and R. L. Riolo. Agent-based modeling vs. equation-based modeling: A case study and user's guide. In *Proceedings of Multi-agent systems and Agent-based Simulation (MABS'98)*, 1998.

[21] S. F. Railsback, B. C. Harvey, R. R. Lamberson, D. E. Lee, N. J. Claasen, and S. Yoshihara. Population-level analysis and validation of an individual-based cut-throat trout model. *Natural Resource Modeling*, 15(1), 2002.

[22] Repast. http://repast.sourceforge.net/.

[23] G. Ropella, S. Railsback, and S. Jackson. Software engineering considerations for individual-based models. *Natural Resource Modeling*, 15(1):5–22, Spring 2002.

[24] R. G. Sargent. Validation and verification of simulation models. In P. A. Farrinton, H. B. Nembhard, D. Sturrock, and G. W. Evans, editors, *Proceedings of the 31st Winter Simulation Conference*, Phoenix, AZ, December 5-8 1999. ACM Press.

[25] T. Schlich. *Molecular modeling and simulation*. Springer, New York, 2002.

[26] N. M. Scully and D. R. S. Lean. The attenuation of ultraviolet radiation in temperate lakes. *Ergeb. Limnol.*, pages 135–144, 1994.

[27] Swarm development group. http://www.swarm.org.

[28] J. Williams. The EPIC model - An overview. In *Natural Resources Modeling Symp.*, 1985.

[29] X. Xiang, G. Madey, Y. Huang, and S. E. Cabaniss. *Environmental Online Communication*, chapter Web Portal and Markup Language for Collaborative Environmental Research. Springer, 2004.

[30] Q. Zhou, P. Maurice, and S. E. Cabaniss. Size fractionation upon adsorption of fulvic acid on goethite: Equilibrium and kinetic studies. *Geochimica et Cosmochimica Acta 65*, pages 803–812, 2001.