

An Agent-Based Watershed Modeling Framework

A Thesis

Presented to
the faculty of the School of Engineering and Applied Science
University of Virginia

in partial fulfillment
of the requirements for the degree

Master of Science

by

Ryan Philip Bobko

May

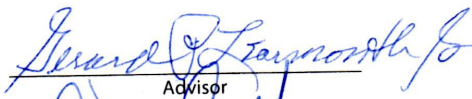
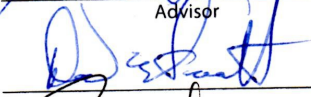
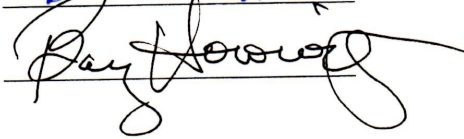
2012

APPROVAL SHEET

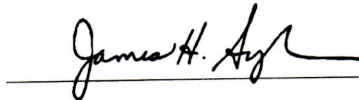
The thesis
is submitted in partial fulfillment of the requirements
for the degree of
Master of Science


AUTHOR

The thesis has been read and approved by the examining committee:


Advisor



Accepted for the School of Engineering and Applied Science:


Dean, School of Engineering and Applied Science

May
2012

An Agent-Based Watershed Modeling Framework

Thesis by
Ryan Bobko

Department of Systems and Information Engineering

University of Virginia
Charlottesville, VA 22902

Thesis Committee:
Gerard P. Learmonth, Sr., Ph.D., Advisor
Barry Horowitz, Ph.D.
David E. Smith, Ph.D.

April 17, 2012

Table of Contents

An Agent-Based Watershed Modeling Framework.....	1
Department of Systems and Information Engineering.....	1
Introduction.....	3
Measuring Watershed Health.....	4
Watershed Modeling.....	8
Nutrients.....	9
Land Development.....	13
Population Growth.....	16
Climate Change.....	17
Existing Models.....	18
Phase 5.3.....	19
Watershed Modeling System 8.4.....	21
BASINS.....	22
Agent-Based Modeling.....	22
Framework Description.....	23
Database.....	28
Flexible Definitions.....	30
Parallelization.....	32
Land Uses.....	33
Agents.....	35
Area.....	35
Household.....	37
River Segment.....	37
Modules.....	38
Geography.....	39
Weather.....	40
Land Development.....	40
Population.....	42
Nutrient.....	43
Policy.....	44
The Engine.....	45
Advantages.....	48
Limitations.....	50
Future Improvements.....	52
Prototype Application – Chesapeake Bay Watershed.....	53
Obtaining and Cleaning Watershed Data.....	53
Methodology.....	57
Prototype Results.....	59
IBM World Community Grid.....	65
Conclusion.....	66

Introduction

A watershed is an area or region that drains into a specific body of water. For example, all water in Chesapeake Bay Watershed drains into the Chesapeake Bay. A watershed can be broken down into sub-watersheds, with each sub-watershed draining into a smaller body of water. Streams and creeks in the James River watershed drain to the James River, which in turn drains to the Chesapeake Bay. This process can be repeated ad nauseum into smaller and smaller watersheds. It is perfectly acceptable to say a person is standing in the Rivanna River watershed, the James River watershed, and the Chesapeake Bay Watershed at the same time. It is important to note that each sub-watershed drains to exactly one larger watershed. For example, the Rivanna River watershed cannot feed both the James and Susquehanna Rivers.

At every level of granularity, watersheds throughout the world are facing unprecedented challenges. Excess nutrient contribution, land development, population growth, and even climate change can affect the state of a watershed. Changes at the watershed level can have wide-ranging effects, including reduced biodiversity, loss of habitat, and decreased water quality. Increased sediment and nutrient loads can lead to hypoxia or anoxia, the decrease or absence of dissolved oxygen in water. Aquatic species can be affected or eradicated in hypoxic regions.

The widespread effects of watershed degradation make modeling especially important, and especially difficult. While a variety of models exist, none account for the possibility of changing human behaviors within a model run. In general, the models currently available follow the usage pattern of setting various input

values and then executing a series of mathematical functions against those parameters. While this strategy has led to a deeper understanding of watershed dynamics, and the results can be compared against empirical data, it may be less effective at forecasting changes during a model run. For example, many models utilize the Hydrologic Simulation Program-Fortran (HSPF), which cannot account for severe weather events. Human behavior may change under different policy decisions, but not in the logical or incremental way many modeling packages assume. There may be a “tipping point” when seemingly small changes rapidly coalesce to produce an unexpected outcome. Nutrients interact in different ways, and are generated from different sources in different geospatial areas. The interactions between disparate pieces of the ecosystem are too difficult to model without gross simplifications, which may then bring into question the results so obtained.

This thesis will present work on a generic agent-based framework for modeling the flow of nutrients through a watershed. The framework was developed to be applicable to any watershed in any condition, and is extensible should that goal not be fully realized. The framework is similar to existing models in its use of functions to describe the state of the watershed, but it adds per-agent specific behaviors that other models lack. A prototype application of the framework was then developed using the Chesapeake Bay Watershed. Results from the framework presented here are compared against the Chesapeake Bay Foundation's modeling efforts.

Measuring Watershed Health

It is not difficult to find an article or news report describing the health of a

watershed. Still, the term “health” is generally ambiguous when applied to inanimate objects. In place of a true definition, various proxies have been developed to describe the state of a watershed.

The Environmental Protection Agency created a “Framework for Assessing and Reporting on Ecological Condition” in 2002 as part of its Healthy Watershed Initiative(EPA, 2002). This framework lists six essential ecological attributes in determining the ecological state of a watershed. However, the areas are quite expansive, and remain difficult

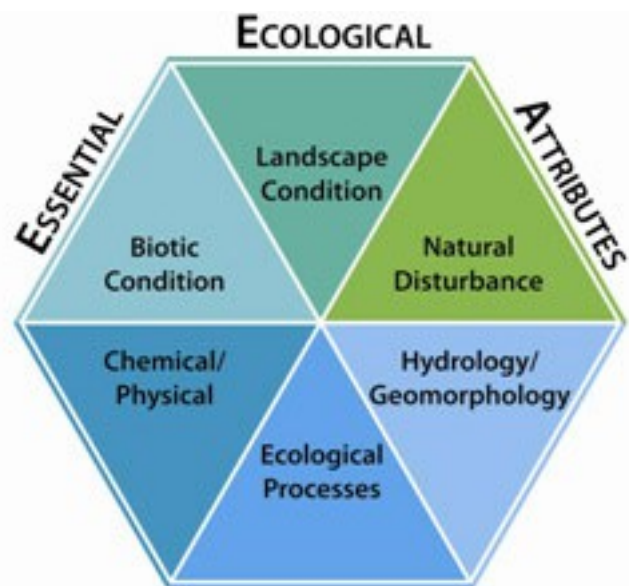


Figure 1: EPA's Ecological Condition Graphic

to measure accurately. Indeed, “Biotic Condition” is one such area in framework. The EPA provides six different guidelines for performing a bioassessment. The guidelines for estuarine and coastal waters has twelve chapters in addition to a section of case studies. Other framework areas are similarly expansive.

Measurement methodologies abound. In practice, they span the range of purely qualitative to statistical analysis.

Qualitative methods can include input from the public about waterway quality, visual stream surveys to identify riparian buffer state and erosion, bird counting, and anecdotal evidence. Booth and Henshaw describe a method of measuring erosion by visual inspection by first nailing markers into the ground or nearby trees, and revisiting the site at on a yearly basis(Booth, 2001). The Ecorse Creek Watershed Management Plan specifies both qualitative and quantitative

factors in determining the status of the watershed(Ecorse, 2001). Qualitative factors include participation of the public in educational programs, stream surveys, and adoption of ordinances by local governments.

The Alliance for the Great Lakes defines several quantitative indicators for measuring the health of sub watersheds in the Lake Michigan watershed(Michigan, 2009). These indicators include items such as the number and density of storm water outlets, impervious surface coverage, slope in watershed, and agricultural coverage. These factors are measured, and then weighted according to a panel of experts. The results were statistically analyzed to determine the health of each sub watershed.

The state of a watershed can be measured quantitatively using various metrics. The EPA has published Total Maximum Daily Load (TMDL) numbers for many watersheds(TMDL, 2012). These numbers reflect the EPA's view on the total amount of pollution a watershed can absorb and still meet federal water quality standards. Pollution in this context may mean dissolved oxygen, sediments, mercury, or nutrients like nitrogen and phosphorus. While the TMDL regulations specify load totals, they do not provide plans or strategies for reducing the amount of pollution discharged into the watershed.

Other sources provide other means of assessing the status of a watershed. The water quality index (WQI) score is one such measure. Though various groups use different tests for determining WQI of a body of water, the score generally measures temperature, pH, fecal coliform bacteria, dissolved oxygen, turbidity, nutrients and sediments(Hallock, 2002). Further, a study in the Las Rozas township of Spain, showed that dissolved oxygen was an easily obtainable proxy

for WQI score in the Guadarrama and Manzanares rivers(Sánchez, 2007). Still, the practice of using a single unit-less score for describing the ecological state of a waterway has drawbacks. For example, a body of water may score well on a WQI scale simply because the scale fails to measure some characteristic in which the water is impaired. There is also an inherent imprecision to a score that is derived from several orthogonal factors. The process of summarizing a number of characteristics into one number obviously entails losing details of the raw data, and thus runs the risk of providing an incomplete picture. WQI's value has been recognized as providing a tool for public consumption and policy makers, but is not adequate for scientific applications (McClelland, 1974).

In general, it seems clear that the spectrum watershed status measurements run from more qualitative in smaller watersheds to more quantitative in larger ones. The size of a watershed may be a limiting factor in its study. This phenomenon reflects the reality that over a very large area, it is impossible to expertly analyze conditions reflecting the state of the watershed. Even quantitatively, determining appropriate numbers for analysis can be a colossal undertaking. For example, the Chesapeake Bay Watershed encompasses 64,000 square miles in six states. It supports more than 3,600 species of plants, fish, and animals, and over 16 million people live in the watershed. The Chesapeake Bay was the first estuary in the nation to be targeted for restoration(CBP, 2009). However, the solution space for this type of inquiry is massive, and direct study over the entire area is impossible. The only alternative to direct study of a watershed the size of the Chesapeake Bay's is modeling. With the advent of computers, such modeling has become a reality.

While there is no standard definition of the “health” of a watershed, there does appear to be some consensus that it is measured against the theoretical pristine state of that watershed. Terms like “biodiversity,” “ecological state,” and even the EPA's “essential attributes” are themselves imprecise terms and concepts in support of the equally ambiguous “health.” Still, it seems clear that plant and animal species are in decline(Levin, 2007), and reversal of this trend can be seen as a “healthy” development. The various measurement rubrics are attempts to measure this progress.

Watershed Modeling

Watershed modeling focuses exclusively on using computer programs to behave *in silico* similarly to how an actual watershed behaves. In practice, this means focusing only on objective metrics that can be estimated or directly obtained. Physical models—building a scale version of a riverbank for example—may be useful for creating estimates or understanding the dynamics present under certain conditions. However, building a scale model of an entire watershed may be as prohibitively difficult as analyzing the watershed itself.

The recursive nature of a watershed feeding a larger watershed leads to complications in terminology as well. For example, there is interest in modeling the Chesapeake Bay Watershed as a whole. However, to model the entire watershed, one must first model the different sub watersheds—Susquehanna, Potomac, Patuxent, Rappahannock, Eastern Shore, York, and James River watersheds. To model those watersheds, one must first model the numerous rivers that feed them, and so on. This recursive terminology affects the

technology greatly. Watershed modeling software often targets a single level of granularity, leaving the results ambiguous in a larger context, or approximations of a finer context. Worse, it may not be clear at the outset what level of granularity a program is designed, leading to wasted effort or dubious results.

The task of modeling watersheds has been approached from many different angles. Some teams have used a mass-balance approach, relying on the principal of conservation of mass to measure nutrient flows(Aschmann, 1999). Others have attempted to define the processes within bodies of water to understand nutrient behavior(HSPF, 1996). Still others combine this nutrient calculus with geographical input and other factors to model the entire environment surrounding the watershed(BASINS, 2007).

The Environmental Protection Agency hosts many watershed modeling projects, though the number and descriptions of them implies that no consensus “best of breed” exists. Several alternative models are described in detail below.

The state of a watershed affects the plants and animals located within that watershed. Decline of a watershed's health has been linked to reduced fish harvesting and reproductive rates. It is seen as a prime driver in the reduction of crab and oyster populations in the Chesapeake Bay, for example(CBF, 2008).

Nutrients

Nutrient measurement is the most prominent objective of the models reviewed. In fact, every model deals with nutrients in some form. Models, such as the Phase 5.3 or BASINS models described below, seek to model nutrient levels in the watershed by modeling changes within rivers as well as contributing factors

on land.

The process of introducing nutrient loads in a body of water is known as eutrophication. Nitrogen and phosphorus are the two nutrients of most eutrophic concern. Together, these nutrients contribute to harmful algal blooms, though the relationship is a complicated one. In some waters, phosphorus is the least abundant macronutrient needed for photosynthesis, and thus limits the growth of photosynthetic organisms. In other waters, nitrogen is the limiting factor. In waterways with high nitrogen levels but in which phosphorus is removed, the nitrogen can be transported further downstream, where it can lead to even larger algal blooms (Anderson, 2002).

Algal blooms can be detrimental to watershed health. A study of Waquoit Bay, Massachusetts found that increased nitrogen loads and concentration led to increased phytoplankton and macroalgae biomass. Conversely, eelgrass cover decreased and was virtually eliminated when nitrogen loads doubled to 30 kg nitrogen per hectare per year (Bowen, 2001). Eelgrass is a seaweed-like plant that produces oxygen and increases dissolved oxygen in a body of water. The algae in algal blooms are also plants, but their lifespan is very short. Blooms lead to increased dead organic matter, which rots and consumes dissolved oxygen. These dead organisms can form rotting hyperscum mats of up to 1m thick. The organisms in blooms can produce hepatotoxins and neurotoxins (Anderson, 2002).

Eutrophication can lead to reduced dissolved oxygen levels, a condition called hypoxia. Perhaps the most visible manifestation of hypoxia is fish kills, where large numbers of dead fish wash ashore or float to the surface of a body of water. Though fish kills can be dramatic on the surface of water, they are no less

likely to occur in its depths. However severe a fish kill may appear, the loss of adult and older juvenile fishes represent only a part of the true effect. Young juveniles or eggs may be even more vulnerable than older fish that exhibit some escape behavior. Also, species with limited movement such as oysters and clams may be particularly harmed. Even when the result is not death, extended or repeated exposure to hypoxic conditions has been shown to slow development growth rates of fishes(Breitburg, 2002).

Nutrient levels in a watershed are generally measured “in-stream,” meaning the amounts are calculated from monitoring stations in the waterway itself. This method is generally considered more accurate than “input-level” calculations, which seek to quantify the amount of nutrients added to land. “Input-level” calculations may not accurately reflect the effects of landscape or stream effects on the nutrients present. For example, fertilizer inputs on land can be over-represented in “input-level” calculations because plants absorb some amount of the available nutrients for growth(Smith, 2000). “In-stream” calculations would not contain this over-representation, but are more difficult to obtain because of the need for active monitoring.

There are numerous sources of nutrients in a watershed. In the United States, agricultural and animal sources represent a significant proportion of the total nutrient load in a watershed. Other sources include point sources, atmospheric deposition, and non-agricultural runoff(Smith, 2000).

Nutrients from agricultural sources are added to soil as fertilizers. Nitrogen and phosphorus are the main ingredients in fertilizer. They increase agricultural production, but plants rarely absorb all the available nutrients during their

lifetime. The remainder is available to be washed off the land and into waterways when the land is irrigated. The amount of nutrients absorbed by plants is known as the uptake value.

Concentrated animal feeding operations (CAFOs) are animal agricultural lands where animals are kept in confined spaces. The US EPA criteria for CAFOs include the number of animals confined and the amount of animal unit equivalents and pollution discharged into waterways. Animal units refer to a standardized way of counting animals of different sizes. For example, a 1,000 pound steer may be one animal unit, whereas a sheep is 0.08 animal units. Feed is brought to the animals in a CAFO instead of allowing the animals to freely graze in pastureland. This concentration of animals and feed also concentrates the deposition of animal waste.

Wastes from CAFOs are managed in a variety of ways, from direct loading into waterways, application of effluent as fertilizer, to storage basins or waste lagoons. These storage basins store wastes in an attempt to reduce the nitrogen content of the swine and cattle-based CAFOs, or phosphorus from poultry-based CAFOs. However, effluent spills or flooding from rains can have an immediate deleterious effect with surface waters(Burkholder, 2007). Even without a waste spill, the storage basins have been shown to contribute to atmospheric nutrient content, which could be deposited elsewhere in the watershed. In addition to nitrogen and phosphorus, waste material from CAFOs can contain heavy metals, pharmaceuticals, and pathogens, all of which can affect the surrounding watershed.

Land Development

Land development is an important factor in determining watershed health. The landscape of the United States has become increasingly urban, a trend that has been replicated throughout the world. The United Nations predicted 51.3% of

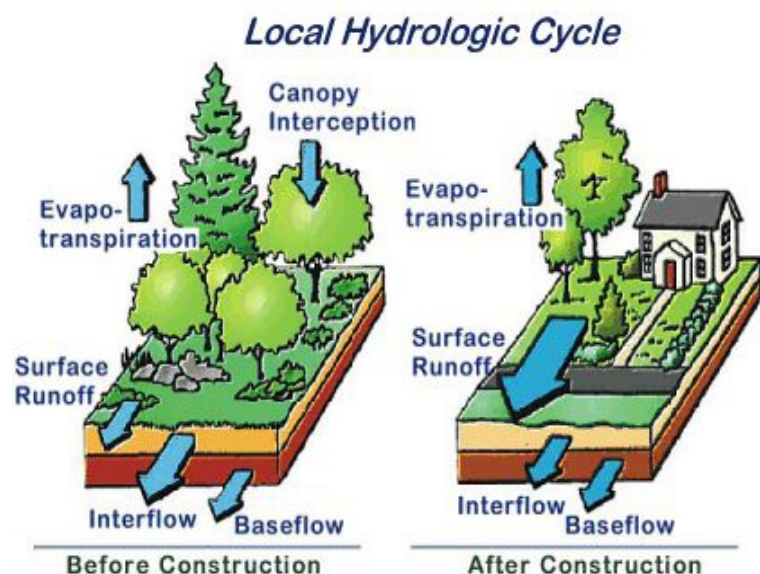
the world population lived in urban areas in 2010. Urban populations in Africa and Asia are expected to double between 2000 and 2030(UNPF, 2007).

Such enormous changes in the character of populations naturally lead to changes in the landscape. In the United States,

Figure 2: Local Hydrologic Cycle

nearly every urban area has increased in area in the last half century, and total urban land area has almost quadrupled(EPA, 2001). The EPA notes that almost all newly developed land was previously farmland, forest, or pasture.

The character of urban land has also changed in the last century. Where cities were once monocentric and compact, they are now often sprawling with suburbs. Population density has also decreased in the largest cities in the United States. The EPA notes that the 34 largest urban areas in the United States have increased area much faster than population. In an extreme example, from 1950 to 1990, Pittsburgh's population increased less than 10% while it's urbanized area grew over 200%. The average ratio of area growth to population growth for the 34



studied areas was over 2.65(EPA, 2001).

The conversion of undeveloped land to urban land can have serious effects on the surrounding watershed. Urban area is generally covered with impervious surfaces like building roofs, asphalt, and cement. This alteration of landscape can increase the volume and rate of surface runoff, and decrease the amount of ground water recharge(Tang, 2005). A study of the Muskegon River watershed's urbanization showed that from 1995-2020 under sprawl estimates, loads for nickel, lead, and oil increased dramatically: 95%, 30%, and 90%, respectively, while nitrogen and phosphorus loads each increased less than 10%. In the same timeframe under non-sprawl conditions, the heavy metal loads were estimated to increase at less than half the sprawl rates; nitrogen and phosphorus increases remained low.

The Muskegon study showed that as sprawl and urbanization replaced agricultural land, nitrogen and phosphorus loads could decrease. It also showed that the effects of urbanization tend to be most prominent in coastal regions of the watershed. The EPA assesses that stormwater runoff is one of the major contributors to ongoing water quality problems in the United States, a problem exacerbated by urbanization. One study showed that a 1-acre parking lot produced over 16 times the runoff volume as an undeveloped meadow, and at faster discharge rates(Schueler, 1995). Furthermore, urban runoff is generally at an increased acidity and higher temperature than runoff from undeveloped land.

Land development is obviously not confined to urbanization. Deforestation and aforestation present environmental concerns as well. Trees store water on their leaves during precipitation (interception), and produce higher rates of

evapotranspiration than shrubs or grasses(La Freenierre). These phenomena slow the release of precipitation into a waterway, which provides a more constant base flow. Without the intermediate processes, rainwater can wash directly into the waterway and out of the area. Much like the effects of urbanization, removing trees increases runoff.

A study of two watersheds in West Virginia showed much the same result(Patric, 1971). The upper half of one watershed was deforested, while the lower half of another was. Both areas were left barren for two years, and completely deforested the following year. Water yields increased after the initial cut, and again after the second cut. Peak flows, water temperature, and turbidity were greater on the lower half deforested watershed.

Urbanization, deforestation, and even agricultural land use can lead to increased sediment load in waterways. Construction sites, regardless of size, have been shown to export significant sediments to the watershed(Owens, 2000). Larger construction sites often have erosion controls to mitigate sediment loss, while smaller sites are less likely to have such controls. In a study of the Nana Kosi watershed in the Himalayas, erosion rates were estimated at 0.09 mm/year for deforested land, and double that for agricultural land. Deforested land erosion rates were themselves over twice that of natural rates. During heaviest rainfall, sediment loads were up to 26 times higher in agricultural areas than forests(Rawat, 1994).

Sedimentation is one of the prime contributors to the decline of aquatic organisms in North America. There are numerous studies showing increased sediment loads and turbidity can increase mortality in many species of fish,

including arctic grayling, rainbow trout, coho salmon, perch, and shad.

Sedimentation may affect the fish directly, as exhibited by reduced gill functioning, but may also decrease reproductive success by reducing dissolved oxygen to the eggs, and reducing spawning habitat(Henley, 2000).

Population Growth

Population growth is very closely tied to urbanization and land development. Though urbanization is increasing disproportionately fast relative to population growth, worldwide populations are increasing. The population of a watershed can adversely affect the watershed's health.

A four-year study of five watersheds in North Carolina showed that fecal coliform abundance was significantly correlated to watershed population (Mallin, 2000). Fecal coliform abundance was also strongly correlated to impervious surface percentage within a watershed, demonstrating the strong relationship between population growth and land development. A study of the Valley Creek watershed in Pennsylvania used a stable nitrogen isotope analysis to measure anthropogenic nitrogen levels from human sewage(Steffy, 2004). Researchers found elevated $\delta^{15}\text{N}$ attributable to human septic systems in all areas below the divide between sewerred and non-sewerred neighborhoods. A stable, rare but naturally occuring nitrogen isotope, $\delta^{15}\text{N}$ has gained traction in measuring nitrogen levels in watershed soil samples, trees, and even humans hair(Robinson, 2001)(Fuller, 2004). While this study provided no means of further identifying the source of the nitrogen, possible sources include defective septic tanks and leaking sewage lines.

The Healthy Harbor Plan for the Baltimore Harbor states that none of the watersheds draining to the Baltimore Harbor meet the State of Maryland's water quality standards for E. Coli. Furthermore, source-tracking studies indicate that human sewage is the main driver for the increased bacterial load. The plan recommends a 95% reduction of human sewage to the harbor to meet state TMDL numbers. Pet waste is responsible for eight to 24% of the bacterial load within the harbor (Harbor, 2012).

Increased human population can also contribute to deforestation. Areas with high population growth and low scores on the human development index (which measures income, health, and education) are correlated to increased deforestation (Jha, 2006). Even areas not experiencing dramatic population growth can see increased harvests of forestland due to non-local demands (Wood, 1998). For example, some of the deforestation of the Amazon basin is driven by population growth, but more is attributed to medium- and large-scale ranchers clearing land for cattle. The demand for cattle is split between local demand of a growing population, and for export.

Climate Change

Global climate change is an area of active study by itself. Much of the study has focused on the increasing levels of carbon dioxide in the atmosphere, and its effects on climate. Numerous studies have reported or forecast increases in global temperatures, more violent and unpredictable weather events. The global average temperature has increased 0.6°C over the past century, a pattern that is expected to continue (Root, 2003). Increased sea surface temperatures and water vapor over oceans suggest an increase in thunderstorm activity (Trenberth, 2005).

This increased convection and water in the atmosphere in turn suggests that when a hurricane forms, it will be more violent and produce greater rainfall. Watersheds will not be spared the effects of global warming.

By one estimate, irrigated acreage will increase and agricultural plant selections will change to account for changes in climate(Adams, 1990). Rising temperatures will also affect watersheds. Hypoxic and anoxic conditions are most severe in summer months, when increased temperatures force dissolved oxygen from the water.

Existing Models

Watershed modeling is not an unexplored area. There are many tools available to a watershed modeler that can model many different aspects of a watershed. Many tools feature an interface which eases “one-off” modeling projects, but may ultimately restrict the utility or reusability of model.

All of the tools described here use the Hydrologic Simulation Program-Fortran (HSPF) as the basis for watershed modeling. HSPF has been in development since the 1960s, and is capable of simulating hydrologic processes from one minute to hundreds of years. It uses meteorologic records to compute water quality results for pervious and impervious land surfaces. HSPF can calculate nearly all variables needed to describe the hydrologic cycle, from interception, evapotranspiration, biochemical oxygen demand, to surface runoff and sediment routing by particle size. The HSPF model contains hundreds of algorithms to achieve its results.

However, HSPF assumes continuous rainfall amounts over the duration of its

simulation period(HSPF, 1996). This assumption leaves it vulnerable to under- or over-reporting actual values in times of extreme weather. For example, a hurricane may significantly increase nutrient levels over a very brief period of time that cannot be modeled with HSPF.

Phase 5.3

The Chesapeake Bay Program's Chesapeake Community Modeling Program has developed the Chesapeake Bay Watershed Phase 5.3 Model of the Chesapeake Bay Watershed. The model has been developed since at least 1982, through multiple major revisions. Each major version has included expanded river segmentation, more land types, and longer simulation periods. It is the model that satisfies the EPA's requirements for calculating nutrient loads in the Chesapeake Bay Watershed for TMDL compliance determination(EPA, 2010#5). Although it is not a general watershed modeling tool, it is included here as the reference model for the Chesapeake Bay Watershed, which is our proposed framework's prototype application.

The system is based on a version of HSPF modified for use in the Chesapeake Bay Watershed. It is augmented by a wealth of auxillary modeling tools used to calculate land development rates, population, and air deposition, and estuarine effects, among other facets. The system is designed to allow smaller watershed models to accumulate into larger areas. For example, individual states in the watershed can manipulate their own data to meet their TMDL goals (EPA, 2010).

The Phase 5.3 model is publicly available as source code (C and Fortran,

with a number of additional processing steps) with instructions for installation. The download, including calibration and data files, is over 800MB. While the target audience is declared to be any watershed modeler, it seems unlikely that a non-developer could successfully install the application. In our attempts, the application could be compiled and calibrated successfully, but never provided any output data. Output is available online, usually in spreadsheets. The system is obviously operational in experienced hands, and its results are essentially authoritative in a regulatory context, but obtaining results is a non-trivial exercise. Phase 5.3 requires a Unix-like operating system, such as Solaris or Linux.

The Phase 5.3 model has drawbacks that make it sub-optimal in several areas. Most notably, the Phase 5.3 model outputs are based on roughly 2000 segments. While this significantly increases the number previously available for modeling, it provides limited insight on the vectors by which pollutants are entering the watershed. That is, the model cannot trace a single small farm's contribution to the health of the Bay, or what conditions are likely to reduce that impact. Similarly, the factors affecting pollution levels in the watershed may have changed significantly in amplitude and character since 2002, the last year included in the model's input data files. Furthermore, the rainfall amounts for any land segment are based on mean daily rainfall over a year, latitude, longitude, and elevation (EPA, 2010#2).

The model employs a number of unorthodox practices when delineating river segments. Some river segments have no size. Other river segments do not follow the documented naming format, or feed into non-existing downstream segments. These discrepancies exist to facilitate calibration with observed values,

or to describe some anomalous situation such as odd river-monitoring station geometries (EPA, 2010#3).

Altering configurations between model runs is a non-trivial task. The system includes over 15,000 configuration and data files, and requires significant effort to recompile and process the files should one change. Altering the model seems outside the reach of any non-expert technician. In fact, this deficiency was noted in 2008 by an advisory committee, which warned of the danger of having only one or two experts knowledgeable in some aspects of the model (Band, 2008).

Watershed Modeling System 8.4

The WMS 8.4 software application is a Windows-based commercial product marketed by Scientific Software Group. It provides a “comprehensive graphical modeling environment for all phases of watershed hydrology and hydraulics (WMS, 2012).” The system relies on GIS overlay computations, and supports a number of hydrological modeling tools such as HSPF, or HEC-1 from the US Army Corps of Engineers. WMS is split into different modules such as a drainage, terrain, and map modules. Each module can be added or removed when using the tool.

The system relies heavily on its graphical user interface (GUI). The GUI is used by the modeler to build models, run simulations, and view results. Models can be defined using GIS overlays, importing aerial images, or reading Computer Aided Design data files. The user can then define stream reaches and watershed terrain, and delineate basins. Results can be visualized in three dimensions. The system has a number of features related to viewing results.

WMS 8.4 appears targeted at smaller sized watersheds. The data entry requirements may be onerous for large areas such as the Chesapeake Bay Watershed. Furthermore, the system's output appears to be GUI-based as well, limiting its utility for running multiple models and comparing results automatically.

BASINS

The Better Assessment Science Integrating point and Nonpoint Sources (BASINS) uses a non-proprietary Geographic Information System (GIS) client to model watersheds. It can use any GIS shapefile or layers. BASINS includes a data extractor, projector, project builder, a number of GIS tools, and decision support tools. It includes an online repository of GIS data and databases for use with the tool, as well as a way to upload new data to the repository. BASINS is a desktop tool for use on Windows computers. BASINS uses either HSPF or PLOAD to model a watershed.

Much like WMS 8.4, BASINS appears targeted at smaller watersheds. The reliance on a user interface limits a user's ability to re-run a particular model to explore the complicated relationship between the pieces of the watershed. The data entry requirements appear mitigated by the online repository, but its performance over a watershed the size of the Chesapeake Bay's may be problematic.

Agent-Based Modeling

Complex systems can be differentiated from merely complicated ones by the phenomenon adaptability and self-organization. Complicated systems—such as a watch or airplane—are designed and constructed from many parts to

accomplish a single goal. Individual pieces may fail and thus bring the overall system to a halt. Conversely, complex systems are able to adapt to changing stimuli. The number of interacting agents is not the defining characteristic; it is the ability to interpret the environment and behave accordingly. This behavior may lead to self-organization among the agents, even though no central organizing principle has been applied (Ottino, 2004). Clearly by this definition, a watershed is a complex system.

Agent-based modeling (ABM) is a technique commonly used to address complex systems. The idea is simple: instead of designing a model that encompasses all known interaction between actors, define the number of simple actors or agents that have rules for interacting with each other. This “bottom-up” approach may lead to previously unknown interactions between disparate agents (Ottino, 2004).

Framework Description

The framework described in this thesis differs from those listed above in a number of ways. It is designed to run unattended, and possibly in parallel. It can be configured programmatically between runs. This flexibility allows a watershed modeler to explore any number of parameter settings without manual manipulation. Perhaps most importantly, it has agent-based characteristics which make it more likely to exhibit emergent behaviors not possible with deterministic or solely equation-based tools (Berry, 2002)(Van Dyke Parunak, 1998).

Furthermore, watershed modeling tools appear to fall into two broad categories: generic ones aimed at small watersheds, or very specific ones

designed and built for a single large watershed. This framework is designed to be scalable from small watersheds to large with the same fidelity at each level. For large watersheds, this framework can free scientists of the need to "reinvent the wheel" for every watershed.

The system is an agent-based modular framework to model non-point sources in a generic watershed. Agent-based systems have a natural analog in software engineering called object-oriented programming (Jennings, 2001). The object-oriented approach to software attempts to create encapsulated pieces of code that can respond to method invocations based on their internal state. Agents in agent based modeling behave similarly—they can operate or not based on their own rules for action. This state can be influenced by environmental factors or neighboring agents. The system is written in C++ to use that language's rich object support.

Architecturally, the system consists of three main pieces: the engine, a set of modules, and a set of agents. Each piece can communicate with the others to perform its part of the model, though in practice, modules generally act as a buffer between agents and the engine. Other components of the framework are more appropriately considered resources for a model run and not part of the



Figure 3: Sample Land Divisions

simulation itself. In particular, the system utilizes a relational database for input, a debugging system for developers, and a variety of filters and aggregators for output. These components are integral to the system's operation and performance, but none directly affects the results of a model run.

The framework calculates its results by dividing the entire watershed into a set of land segments. Each land segment can be further divided into one or more parcels. Each parcel is then divided into areas. Waterways are similarly organized into smaller and smaller divisions, from watershed (the entire river system) to basin to river segment. Together these features account for all land and water in the watershed.

Land segments have a defined size in acres that does not change for the life of a model run. Otherwise, there is no restriction on the definition of land segments in the framework. Some situations may call for political delineations instead of hydrological ones, so each land segment can have some percentage of its land area in the watershed. Each parcel in a land segment can likewise occupy some percentage of the land segment's area. Each parcel contributes some percentage of its total nutrient load to each intersecting river segment.

The system runs from a user-specified start date for a user-specified number of steps, or ticks. A tick is a discrete unit of time in which a set of related calculations are grouped. The system's tick length is defined as a one-month. Thus, a twenty-year model run can be performed in 240 ticks. The selection of a one-month tick period has several rationales:

1. For the prototype application of the Chesapeake Bay Watershed, much of

the required data is available only in monthly form. Other data is available as yearly data.

2. While weather data is available at a higher granularity, interrelated activities like fertilizer applications are estimated even at the monthly timescale.
3. Monthly numbers provide a reasonable trade-off between model fidelity and computational requirements.
4. Output numbers can be easily converted between tick number and year/month values, for reporting purposes. Longer periods, such as seasonal changes, can thus be aggregated. However, the reverse operation on a longer tick period would introduce ambiguity in the reporting.

The output system is user-configurable to provide multiple aggregation levels. By default, the output is unfiltered and consists of nutrient levels per tick per agent. The output consists of:

- Tick number
- Agent ID
- Location of the agent (Area)
- Nutrient
- Value
- Source

This output is the most verbose, and requires the most disk space. Slightly

more concise output can be retrieved by aggregating based on areas instead of agents. The system also supports aggregating by parcel, or only issuing results at the end of a tick, with no raw results posted. In this mode, the system provides end of tick nutrient accumulations and their sources. Finally, the system can generate descriptive statistics per tick per land segment instead of raw data. This aggregation level is approximately as verbose as the default option, but it does not provide any access to the raw data. Statistics provided are:

- Number of values computed
- Minimum value
- Maximum value
- Mean value
- Median value
- Mode
- Standard deviation

The framework also provides a debugging system, which produces results similar to the output system. However, the debugging system can trap logical errors in setup such as inconsistent input data. It can provide extremely detailed information about how the values sent to the output stream are being calculated, such as how population growth is occurring within an area, or how acreage is changing. While similar to the model output, such information is often more useful during model development than during execution, so the streams of data are separated for convenience.

Database

The cornerstone of this framework is a relational database describing the watershed being modeled. Relational database management systems (also known as RDBMSes) were first introduced by Edgar F. Codd in 1969(Codd, 1969). Since that time, there have been numerous commercial and open source implementations including Oracle, Microsoft SQL Server, and MySQL. Until very recently, the relational database has been a ubiquitous component in any large data retrieval system, and it still dominates structured data retrieval systems in almost all settings. So ubiquitous has been the use of RDBMSes that the generic term “database” has come to mean RDBMS, though other database systems do exist(OODBMS, 2012)(NoSQL, 2012).

One strength of relational database systems is the use of the Structured Query Language (SQL) to select, update, delete, and insert data. SQL has largely been standardized, and every major database provider supports this standard to one degree or another(Date, 1997). Even niche providers support the standard, providing near universal portability between different systems. While small differences in syntax do exist, the concepts of the language are consistent among all RDBMS providers.

SQL's model of data is one of a “schema” containing “tables” representing one type of data. Each table is composed of attributes called “columns.” The data is then organized as “rows” in a table, meaning each row will have a value for each column in the table. Tables can be linked together, providing a means of relating data in one table to data in another. SQL provides the language for doing set operations on these tables.

Figure 4 shows a simple relational database schema linking states to land segments, land segments to parcels, parcels to river segments, and river segments to river basins. From this schema, it is possible to retrieve all parcels in Virginia using the following SQL.

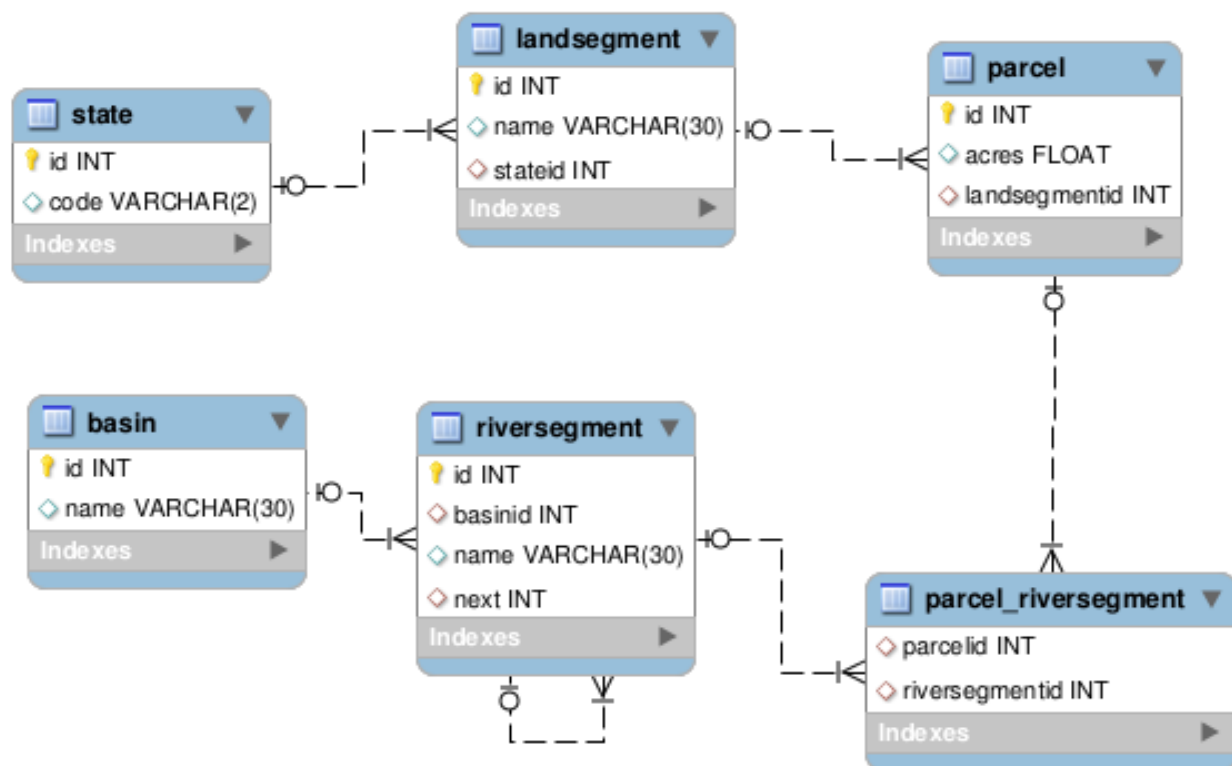


Figure 4: Database Schema

```

SELECT parcel.id FROM parcel JOIN landsegment ON
parcel.landsegmentid=landsegment.id JOIN state ON
landsegment.stateid=state.id WHERE state.code='VA'
  
```

A slightly more complicated statement could retrieve the basins with river segments in Virginia. Note that this statement does not define the number of parcels expected, but only what parcels are desired and how to link them together. Thus, there could be millions of river segments or land segments rows in the database, and the user would operate on a set in the same fashion, regardless of how many rows are returned.

This framework seeks to harness the tremendous power of the underlying

database system to model a generic watershed. The hypothesis is that a sufficiently general database schema is capable of modeling any watershed. Restated, what differentiates one watershed from another is the actual data and functions describing the watershed, not the relationship between different pieces of data. This generality can be achieved through proper definition and categorization of stresses affecting a watershed, its structure, and the relationship between its disparate pieces.

Many aspects of the framework's data model are influenced by the flexibility imparted by the database schema. In particular, functions, nutrients, and best management practices have no *a priori* definition in the framework. Rather, the framework employs the concepts of functions, nutrients, BMPs, with the database supplying the definitions.

Flexible Definitions

Perhaps the most basic element of the framework is the concept of a nutrient contribution function. Each agent's nutrient contribution is defined as the result of a function and the type of agent. More generally, the system works largely by calculating results of functions and applying the results to the question at hand. This is a similar approach to other watershed modeling tools, described above. Its accuracy as a modeling tool, then, is dependent on how accurately its functions mimic the real world behavior they are modeling. To that end, the framework includes a number of function types. Available function types are:

1. Constant: $f(x) = \text{Value}$
2. Linear: $f(x) = mx + b$

3. Limits: $\text{Low value} \leq f(x) \leq \text{High value}$
4. Exponential: $f(x) = Ar^x$
5. Logarithmic: $f(x) = a + \ln(x)b$
6. Stepped: Stepped functions represent different function types across a range of possible inputs. This meta-function provides for discontinuous values as the input changes. For example, a population growth function could be linear until some date, after which it becomes exponential. Up to five discontinuities are supported.
7. Composite: Composite functions generate output by successively applying a list functions to the input value. The output of one function becomes the input of the next.

All land area, population, and nutrient contribution functions use the current tick as the input value. This imposes no time constraints to the model's run, provided the functions are defined for every tick. Of particular note, very large (small) exponential functions could result in a over-(under-)flow of the computer's double-precision capabilities. Trapping this condition is an area in need of further study.

The framework makes no assumptions about what nutrients will affect a given watershed. For example, the Chesapeake Bay Watershed is largely affected by excess nitrogen and phosphorus, while the Athabasca River is affected arsenic, mercury, oil, and bitumen from oil sands mining runoff(EPA, 2010)(Hrudley, 2010). With such a diversity of elements to be measured, there is no way to pre-load the system with every possible nutrient. Instead, the system allows a user to define

nutrients by adding rows to the *nutrients* table. Nutrient contribution functions can then be defined for each agent or area.

In this light, a nutrient is simply a measurable quantity of something, which agents contribute to the watershed. A watershed may be under stress from added mercury or iron ore. Sediments may be treated as a nutrient, as may be dissolved oxygen. This extremely loose definition gives enormous flexibility to the watershed modeler.

This strategy is also employed with respect to best management practices. There is no limit to the number or effect of BMPs that can be included in the framework. For reference, the prototype application modeled 23 distinct BMPs. Each BMP can be associated with one or more land usages, though the BMPs effect on nutrient contributions are determined by the area that employs it. This situation accurately models real world environments, where some BMPs are more effective than others, based on geography, environmental conditions, or merely expertise(Washington, 2008)(Park, 2007). Each BMP can affect one or more nutrients with different efficiencies.

Parallelization

The problem of watershed modeling is an expansive one. The computing power available to researchers has come into wide use in the last several years, and opened a new set of problems to computational exploration(Ekman, 2004). The framework described here can run thousands of calculations per tick in only a few minutes. Still, even this level of performance may not be enough for very large watershed models. A few minutes per tick over hundreds of ticks can be

several hours of compute time.

Though computer performance is expected to continue to improve (Ekman, 2004), the framework can operate in parallel to reduce runtime. In particular, the recursive structure of watersheds lends itself to data parallel execution. Because watersheds cannot overlap, the nutrient contribution of each sub-watershed can be calculated independently of every other sub-watershed, with the values accumulated afterward. This parallel execution is clearly not perfectly parallelizable because watersheds may be different sizes and take different amounts of time to execute. That is, if the watershed in question had two sub-watersheds, one much more complex than the other, one compute node might finish sooner than the other. With equally complex sub-watersheds, the computation would be expected to be more perfect.

Furthermore, the task of calculating response surfaces around a set of parameters is embarrassingly parallel. That is, multiple models of a unique parameter set can be executed on many machines at the same time. The results can then be analyzed to discover the parameters most responsible for nutrient flows in the watershed. This strategy was employed during the prototype application investigation, and expanded for use on the IBM World Community Grid.

Land Uses

In the real world, every piece of land in a watershed is used for something. As described above, this usage can affect its nutrient contribution, and which BMPs are available to it. For example, a farm is more likely than a mine to use

fertilizer, but the farm may benefit from a strategy of no-till agriculture, which makes no sense for the mine.

The framework accurately models the real world situation. Each area in the framework has a land usage that affects its behavior. Furthermore, land usage defines a population density that determines how population change is allocated to areas of that type. All contribution and population numbers are configurable in the framework.

The available land uses are:

1. Urban and Suburban: Urban and suburban are represented in the framework by two distinct land use types to allow them to have different population densities. However, in other aspects, the two types behave in the same way. Each area can be either high- or low-density, and represent impervious or semi-pervious land cover.
2. Farm: The farm usage type can represent both animal and crop farms, but also undeveloped meadow or pasture. Each farm has crop or animal type, a tillage strategy, does or doesn't add manure, and has some sort of nutrient management plan. Furthermore, this land type has a "bad buffer ratio" to represent degraded riparian buffers.
3. Centralized Animal Feeding Operation: CAFOs contributions generally have inflated nutrient contribution functions, described below.
4. Construction and Mining: Much like CAFOs, construction and mining areas can have a different nutrient profile, depending on user-supplied data. They can also have different population densities, though in the real world,

construction areas become urban or suburban areas. As such, construction and mining areas generally have no population.

5. Forest: The forest usage type accounts for forests and harvested forests, including clear-cut areas.
6. Water: This usage type generally has no population and would rarely change acreage, but can still affect nutrient contributions through its contribution functions. It includes real world usages like lakes and reservoirs.

Note that all land uses are can export nutrients to the watershed, as controlled by their nutrient contribution functions. In that sense, the land usage types themselves are a shortcut for behavior that could be modeled using more descriptive nutrient contribution functions.

Agents

Agents are the fundamental building block of any agent-based model. The framework defines certain classes of agents and their attributes, but enforces no limit on their number. For reference, the prototype application described below involved upwards of 30,000 independent agents. This number of agents provides a workable balance of detail and aggregation to our model(Doran, 2006). To facilitate the possibly enormous number of agents required for a model run, the agent data must be organized in a scalable way.

Area

An area is the most common type of agent in the framework. Each area consists of an amount of acreage, a land use, a population situated in some number of households, and a set of nutrient contribution functions. An area may

have a different nutrient contribution function for each month of the year. For example, a farmer may choose not to fertilize cropland in wintertime. An area also has an edge-of-field (EOF) ratio to simulate the transfer of nutrients from an area to its neighboring river segments. Each area can have its own EOF multiplier function for each nutrient. These functions calculate EOF ratios from the flow of the river segment.

Within each tick, every area agent is queried for its nutrient contribution and that contribution's source. Internally, the agent calculates this number by delegating the question to its households and adding the results of its own nutrient contribution function, if any. An area's nutrient contribution source is determined by its land usage. The framework supports multiple sources for an agent's nutrient contribution, however.

Every area has the possibility of employing zero or more BMPs that modify its total nutrient contribution. Further, an area may employ a BMP over only a subset of its total acreage. For example, a urban area may have separate stormwater sewage systems covering 25% of its acreage. Farms may use continuous no-till practices on a portion of their fields. In determining total nutrient contribution, the area calculates the effect of every BMP in use by first calculating the efficiency of the BMP in that area, the acreage covered by the BMP, and which nutrients are affected by it, and scales the total contribution accordingly.

$$Nutrient\ load = EOF \times Land\ segment\ percentage \times \sum (Contribution - BMP \times efficiency \times coverage)$$

Household

A household is a simple agent that contains a population and some acreage of lawn. The framework supports household types of apartment, townhouse, and single-family house. More acres in the United States are covered by lawns than corn. (Milesi, 2005). Studies have estimated that lawn fertilization can have a significant nutrient export to the watershed. During summer months, each household can decide to fertilize their lawn, or not. The framework assigns randomized fertilization frequencies from zero to four times per year. Once determined, a household will always fertilize with the same frequency every year.

The inhabitants of a household are assumed to be humans, and create an amount of human waste every month. This sewage—biological, from household cleaners, or other sources—is handled via either a septic tank or a sewer system, depending on the land usage where the household is located. Each area can have a different ratio of septic/sewer systems, and can specify different per-person nutrient loads. Septic contributions are not subject to EOF scaling.

River Segment

A river segment is the smallest unit of waterway in the framework. Each river segment feeds exactly one downstream river segment. The amount of water that flows from one river segment to the next is represented in the framework by a flow variable, which can be scripted for every month and year during a simulation. The river segment agents contain a set of flow multiplier functions that simulate how a given flow affects “in-stream” processes such as deposition, scour, or denitrification. The flow multiplier function can be different for each nutrient in each river segment.

The flow multiplier is effectively a discount factor for nutrients, a situation that occurs in the real world. For example, Virginia regulations state that nutrients added to the Chesapeake Bay Watershed can be traded among regulated entities. However, the trade is weighted based on the utilities' locations in the watershed (VPDES, 2012).

Modules

Modules provide the basic computing organization of the framework. Modules can calculate exogenous values for use by agents, as well as spawn their own agents to be part of a model run.

Every module implements an interface for interacting with other modules and the framework engine. This interface has six distinct phases:

1. **Configure:** The configure phase is the first initialization opportunity for the module. During the configure step, the module can initialize itself to an internally consistent state without communicating with other modules.
2. **Setup:** Setup is the second initialization opportunity for a module. In the setup phase, communication between modules is possible. A module may need information available from other modules to prepare for the ticks to start. For example, the population module may need information about land development during initialization.
3. **Prepare:** The prepare phase is executed before every tick, and provides the module an opportunity, for example, to read the previous tick's results prior to the next tick starting.
4. **Run:** Modules that spawn agents will generally advance their agents by one

tick in this phase. Not all modules must spawn agents, however. Modules not responsible for agents would generally calculate any values needed in the prepare phase.

5. Finish: Execute any post-tick cleanup, or values that should be read during the next prepare phase.
6. Checkpoint: Write any checkpoint information to stable storage. This phase allows a module to be restarted if the system fails before all the ticks of a model run have been completed.
7. Teardown: Do any cleanup necessary before the model run ends.

Geography

The geography module organizes land and river segments for a model run. It creates every land segment and its associated parcels, as well as all river segments. It links river segments to their downstream river segment. Furthermore, the geography module maintains the record of how much of a land segment is in the watershed, and what percentage of a parcel's nutrient contribution should be applied to a specific river segment.

Beyond the basic module interface, the geography module can provide generic geographical information to other modules and/or agents. This information includes:

- Which river segments are present in a given parcel
- Which parcels supply nutrients to a given river segment
- Which river segments immediately feed a particular river segment

- An ordered list of all river segments that feed a particular river segment

The framework assumes there is exactly one river segment that is the root of the river network. This root is the ultimate downstream river segment for all river segments in the system. For example, in the prototype application, the Chesapeake Bay itself is the root of the river network. This requirement makes possible the orderly traversal of all other river segments in the model.

Weather

The weather module provides exogenous weather and ecological effects to the watershed. It provides plant uptake values for each nutrient and area in the watershed for each tick. It also updates river segment flow numbers to reflect increasing or decreasing flow for that tick. River segment flows can be scripted per month and year for each river segment. This capability provides a means to script weather events throughout a model run. It is an example of a simple module that can have widespread effects on the agents, and the model results as a whole.

Land Development

The land development module controls the conversion of land usages in the land segments. During its configure phase, the module spawns all area agents for simulation. Areas exist for the duration of the model run, though they are expected to change size repeatedly.

Because the framework does not impose a starting year, the acreage of each area cannot be defined beforehand. Instead, each area uses development trajectory function to calculate its acreage. This trajectory function determines

whether the area expands or contracts during a model tick. The initial size of an area is therefore calculated by executing the function for the first tick. The system sets the minimum size for an area at 0.001 acres, or about 44 ft².

There is no limit to the area trajectory functions. Area calculations are therefore sanity-checked against the total acreage in the land segment, and modified up or down in case of a disparity. All areas with non-constant trajectory functions are uniformly scaled to eliminate the difference.

All area types except farms are created in an identical fashion. The acreage is calculated, and a single area is created to occupy that space. This means that every parcel will have exactly one agent of any given land usage. An alternate strategy is employed when creating farm areas. With farms, the total acreage is calculated as usual, but a “meta-farm” is created to occupy that space. The “meta-farm” initializes itself with any number of smaller independent farms. The distribution of farm sizes within a given acreage is determined using a stepped function with the available acreage as the input value. The farms that compose a “meta-farm” behave independently in every respect. The extra level of abstraction provides a convenient way to allocate the potentially enormous number of farms in a watershed. It also provides a way to differentiate a wide variety of farm sizes present. For example, a 10,000 acre industrial farm is likely to behave differently than a 10 acre backyard farm.

Other modules and/or agents may need other information from the land development module. In that case, it can provide a list of areas for a parcel, and a list of areas that have a specific land usage.

Population

The population module is closely related to the land development module. While the land development module calculates acreage changes for all area agents in the model, the population module assigns inhabitants to them. Every area with a non-zero population density must contain at least one inhabitant. Much like the land development module, the population module cannot determine population from the outset, but must calculate population based on a population trajectory function for each land segment. This trajectory is modified by the percentage of the land segment in the watershed. This feature accounts for the possibility that a land segment may not be totally inside a watershed's boundaries, but no watershed-specific population numbers are only available for the land segment. In the United States, population numbers are generally available on a per-county basis. When a land segment contained more than one parcel, population was apportioned based on parcel acreage.

Once the total population for a parcel is calculated, the total acreage weighted by population density of the land usage types is calculated for that parcel. Total population is assigned to the parcel's areas based on an area's relative size and population density. Once the population is determined for each area, it is allocated to households.

When the population of an area is increasing, new households are created to absorb the increase based on watershed-wide household median size and standard deviation. Decreasing population is handled in roughly the same manner: households are removed until the population decrease is satisfied, as long as at least one household with at least one inhabitant remains in the area.

The population module/land development interaction provides a perfect rationale for multiple initialization phases in the model lifecycle. The land development module is capable of determining area acreages without input from other modules, but the population module cannot assign population without the land development module publishing its acreages first.

Nutrient

The nutrient module is the heart of the simulation. It is responsible for bringing together the disparate pieces of information provided by the agents and other modules, and determining the nutrient load for every model tick.

The bulk of the calculations are performed while looping through all the areas in the simulation. Each area is queried to determine its nutrient contribution for each nutrient. This number is then scaled to account for that area's EOF ratio. The scaled result is further scaled to account for the amount of the containing land segment is in the watershed.

Once the nutrient contributions are calculated for each area, the values are accumulated to the parcel level. The nutrient module then traverses the river segment network starting with the root river segment. At each level, a river segment's total nutrient load is determined from the parcels that contribute nutrients to it (scaled to reflect a parcel's ability to feed multiple river segments), plus the contribution of upstream river segments.

$$\text{Nutrient load for river segment } R_i, N_{R_i} = \sum N_{R_{i-1}} + \sum_{\text{abutting parcels}} P \times \text{percentage}_{p, R_i} \times \text{Flow}_{R_i}$$

The recursive nature of this algorithm ensures each river segment is visited exactly once, and that nutrients propagate from the farthest reaches of the river

system to the root of the network.

Policy

The policy module is designed to allow a modeler flexibility in configuring a model run without directly changing the database schema. These policies provide a policy maker the means of exploring “what-if” scenarios for a model run. The framework contains two basic types: policies and sliders. Policies are temporal in nature, and can change throughout the course of a model run. Sliders are constant throughout a model run.

Policies can be defined for areas or parcels. A policy specifies a year, month, nutrient, and an amount. Each policy can define what the amount value means. The framework currently provides two policies: gross maximum nutrient contribution and percentage nutrient contribution. If defined, these policies are applied during the nutrient module calculations after the total nutrient contribution is calculated. Extended, policies could affect any aspect of the model.

Because policies can be defined for every model tick, they provide a convenient method for scripting policies throughout a model run. For example, a policy can affect some aspect of the simulation for some specific time range.

Sliders are constant throughout the model run, and are defined for areas. There is no limit to the number of sliders that can be defined for an area. A slider specifies the area, nutrient, and amount, though the nutrient is optional. Slider values are used heavily during the model initialization, though they could be used at any point of a model run. The framework defines nine sliders:

1. Population septic nutrient contribution in pounds per person per month

2. Population sewer nutrient contribution in pounds per person per month
3. Fertilizer per application in pounds per acre
4. Median household size
5. Standard deviation of household size
6. Percent lawn coverage
7. Percentage of population on a sewer system
8. Population rate scaling factor
9. Development rate scaling factor

The Engine

The primary functions of the engine are to initialize the other components when a model run is started, coordinate their operation during a run, and to ensure atomic operations for agents. The engine is effectively a bookkeeping apparatus, with no direct affect on the model's results. Such an approach has been employed successfully in other large-scale ABMs (Carly, 2006).

The engine provides atomic operations on agents to ensure the scheduling algorithm does not introduce bias to the results. Update policy can have a dramatic impact on the resulting calculations. Haphazard updating can lead to model bias that may affect the study's conclusions (Schonfisch, 1999). Different ABM frameworks rely on different update strategies. Netlogo relies on a "turn-based" strategy, where sets of agents divide their movement and calculations equally. Repast feature a serial update strategy, with an optional randomized starting location. Swarm effectively maintains two copies of the data: one for

calculating updates, and one for writing them (Welch, 2010).

The number of agents this framework can support makes multiple copies of data impractical. Instead, read consistency is accomplished via a temporary storage journal that agents can register values to be applied later. During the course of a model run, an agent may calculate some next state that it will have. To ensure consistency between its responses to any previous query and future queries in the same tick, it cannot immediately convert to this next state. The engine's agent journal provides a means for the agent to store for later use any data it will need. Should the computer running the simulation fail before completing its calculations (for example, in a grid environment), the tick can be restarted without replaying the journal and without loss of data.

The agent defines what information is stored in the journal, and is responsible for interpreting the data sent by the engine. The engine provides the data back to the agent during pre-tick preparation. An agent may have any number of journal entries per tick. It is assumed that a relatively small percentage of agents will make use of the journal in every tick.

Initialization includes such tasks as identifying and initializing the input database, setting the output aggregation system and debugging granularity. It also includes identifying the modules to execute during a run. It sets various configurable properties, such as the length of the simulation run (number of ticks), the starting year, and an optional random seed. When a random seed is omitted, the system's start time is used to seed the random number generator. The engine keeps track of the current tick, and provides tick-to-date and date-to-tick utilities.

The engine can apply arbitrary SQL statements to the input database during initialization. The database is loaded from disk into memory, and the SQL is executed against the in-memory version. This allows for multiple model runs to be performed with different data without requiring multiple databases on disk. When executing thousands of model runs, this capability can save a significant amount of disk space.

After the first level of initialization is complete, the system moves the specified modules through their configure phase. Once all modules have completed their configure phase, the engine starts the modules' setup phase.

After module initialization is complete, the engine goes into a loop for each tick. The loop includes:

1. Start the prepare phase for every module
2. Afford all registered agents the opportunity to set values for the coming tick
3. Commit any values back to the agents
4. Start the run phase for every module
5. Start the finish phase for every module
6. Start the checkpoint phase for every module
7. Flush the output aggregator

Once this loop is completed for every tick, the engine begins its cleanup operation. This entails flushing the output aggregator a final time and putting all modules in the teardown phase.

Advantages

The framework has a number of advantages over existing models. Notably, the use of agents in the framework can lead to unexpected outcomes that are not possible in purely function-based applications. The results of this framework could be unexpected, and reflect actual interactions of different agents in the watershed under different stimuli. The agents can react to their simulated environment in a way that is not possible with non-agent-based models.

The framework described here is capable of modeling extreme weather events via its flow multiplier and EOF variables. Extreme weather events are considered essential in understanding watershed dynamics (Brezonik, 2001). Times of extreme weather—hurricane activity, torrential rains, thunderstorms—can greatly increase the nutrient load in a watershed. The increased water can overflow waste storage basins and stormwater systems, increase soil erosion rates and wash more surface nutrients into a waterway.

The framework was designed for portability and extensibility. Considerable effort has been expended to reduce the framework's runtime memory and disk footprint. It can be compiled and run on a variety of operating systems, including Windows, Linux, and Unix-like systems. There are several extensibility points that provide watershed modelers access to changing the model behavior. Extending the model requires knowledge of C++ programming methodology and possibly SQL, depending on the desired changes. The framework contains many fully-implemented examples of agents and modules, providing a guide for the development of replacement pieces.

The primary extensibility point is through modules. A new module could be

implemented by designing a single function, the run phase. The parent class supporting all modules is fully implemented except for this function, meaning no extra code is required. If an updated module is required, even this is unnecessary if not needed. For example, a replacement weather module could be developed to more accurately provide weather information. Such a module would require reimplementing only the necessary phases, and using the existing weather module's phase implementations for the rest of the module operation.

New land usages could be developed by adding new agent code, and possibly registering the new type in the development module. For example, changing all land development allocations to use the farm-size allocator would not require any specific land development module changes, but would require modifying the agent initialization code. This could be accomplished for a single land usage type, or all types, as needed.

Similarly, new function types, policies, and sliders could be introduced by inserting the appropriate rows to the database, and providing a software implementation to interpret these new values. It should be noted that the database rows would not require changing the schema, but only adding new rows to the existing tables. This practice is no different than what a modeler would do as a matter of course when modeling a new watershed.

Because of the general nature of the database schema and extensibility designed into the software, the framework can support any watershed. In the general case, a watershed's various nutrient contribution, land development, and population functions, as well as the particular nutrients to be modeled, can be loaded without requiring any new software development. Furthermore, runtime

values can be changed between model runs using industry standard SQL, greatly reducing the chance for human error. For example, modifying runs of the Phase 5.3 model requires editing a possibly large number of data files and a recompilation of the source code.

This capability is markedly different from other large watershed modeling applications. For example, the Phase 5.3 model is used to model the Chesapeake Bay Watershed, but provides no support for modeling any other watershed. It is our belief that no particular requirement or restriction prevents its use in other watersheds, other than narrowness of design ambition. This narrowness is perhaps understandable when one considers that the Chesapeake Bay Foundation's mission statement includes no language about general modeling (CBF, 2012).

The framework can run unattended on one or more machines at the same time, and is restartable. The agent journal provides for consistent data retrieval even in the case of a system failure. In addition, the database schema and data layout is conducive to parallel computation, a feature almost no other modeling applications contain. In fact, the Phase 5.3 Model makes no parallelization claims, and the desktop-based tools reviewed are confined to the host computer.

Limitations

The framework was developed to deconstruct watershed modeling from one large monolithic program to a small framework architecture with a large number of simple agents. It is our belief that this paradigm is ultimately more likely to contribute insights to watershed dynamics, regardless of the size and complexity

of any given watershed. However, the system does have limitations that must be addressed.

The framework contains a simplified “in-stream” nutrient model. That is, once nutrients enter a waterway, they are merely transported to the root of the river network with one function describing the journey. In reality, nutrients continue to change while en route to their final destination. Such changes include nitrogen and phosphorus cycling, deposition and scour, and denitrification. Transport simplicity fails to provide a precise picture of nutrient flows in a watershed. Additionally, a river segment's ability to carry nutrients is more properly represented via a channel rating curve or capacity variable, possibly based on cubic feet of water per second and some representation of the channel dimensions. The flow multiplier function is less expressive than necessary.

The framework provides an adequate selection of agents, but few agent-specific behaviors. However, updating agent behaviors requires writing software extensions to the model. A truly generic solution would not provide agent classes such as “farm,” “urban,” or “construction,” but would instead provide a means for the user to formulate agent classes relevant to watershed being modeled. Likewise, waterways could be extended to include not just river segments but also ground water, evapotranspiration, and other qualities.

Agent-specific behaviors are likely to be markedly different among watersheds. The agents provided in this framework are perhaps too simple, and do not change behavior or react to their environment enough to effect widespread changes. Creating an extensible way for a watershed modeler to imbue the agents with unique behaviors given a set of circumstances remains an elusive

goal.

Future Improvements

There are several improvements possible in the framework, including the limitations listed above. Module improvements are an area of great potential. The weather module is little more than a random number generator when it could be much more realistic. It could model sun and precipitation to arrive at the projected uptake values for each plant. This approach could use precipitation values to drive the river segment flows. Different regions in the model could have roughly the same weather conditions. For example, the Chesapeake Bay Watershed is very large, and New York may be experiencing a cold wave while Virginia is undergoing a heat wave. In particular, a weather module that can occasionally generate extreme storms would be very beneficial.

The type of modules needs to be expanded. The current framework lacks an air deposition module to simulate airborne nutrients landing in the watershed. Air deposition could be informed by another new module: an economic module. For example, a growing economy could lead to higher energy requirements, which would lead to increased airborne mercury or particulates.

Agents in the framework are autonomous, and can behave based on their own set of rules during a run. However, this autonomy is rarely seen in the real world. In reality, agents are often very interconnected, and can influence each other's behavior. The framework must account for some agent autonomy as well as agent interactions.

The framework makes limited use of geographical data, which is not

optimal. Geography is handled as a byproduct of the tree-traversal of the river segment network when calculating nutrient loadings. In this manner, a land segment's distance from the root of the river network can only be given in river segments “hops” with no sense of distance. In reality, river segments are not of uniform size, and calculating distance from the root river segment may be of some value.

The geographical aspect of the system is more important when determining runoff values of an area. An area that is very distant from a river but still in a river segment would likely export fewer nutrients to the watershed than an area bordering a river.

Prototype Application - Chesapeake Bay Watershed

To validate the effectiveness of the framework, it was used to model the Chesapeake Bay Watershed. The Chesapeake Bay is the largest estuary in North America. Its watershed is composed of parts of six states and covers 64,000 square miles. The Bay has been under stress from population growth, nutrient pollution, and deforestation for decades. The results of these stresses include reduced submerged aquatic vegetation, degraded wetlands, and declining oyster, crab, and fish populations. The Chesapeake Bay Program was created in 1983 to address these issues, with amendments following 1987, 1992, 1994, and 2000(CBP, 2009). However, the health of the Bay and how to improve it is an open question.

Obtaining and Cleaning Watershed Data

The Chesapeake Bay Watershed has been studied for decades. This body of

work provides a wealth of data about the status of the watershed. The data for the prototype runs was drawn from several empirical sources. Much of the data is available via geospatially-unique Federal Information Processing Standard (FIPS) codes, which generally comprise exactly one county, but may be assigned to population centers not enclosed in counties (such as Charlottesville, Virginia). The population model was created from US Census estimates from 1980-2009 at the FIPS code level, and extrapolated into the future. Land usages and conversion rates are based on the Chesapeake Bay Program's Phase 5.3 model dataset. Nutrient contribution numbers will be based on the Phase 5.3 model dataset. Septic system contributions and coverage were estimated from various sources. Representative farm sizes were calculated from data from the National Agricultural Statistics Center.

The process of cleaning the data can be onerous. The framework utilizes functions for a wide variety of calculations during a model run. A separate application was developed to marshal the existing data files and calculate the needed regressions. A limit of 0.8 was

used as a discriminant for the coefficient of determination (R^2) when choosing an appropriate regression.

The following algorithm was employed to fit regressions to the data:

1. If all values were identical, or a single number existed, use a Constant function.

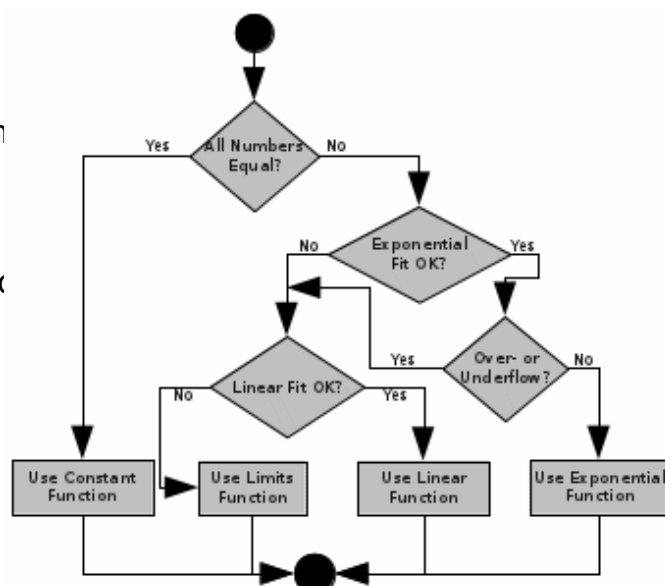


Figure 5: Regression Decision Tree

2. If the exponential fit is better than the minimum R^2 and better than the linear fit, and there is no over or underflow for possible model run times, use an Exponential function.
3. If a linear fit is better than minimum R^2 , use a Linear function.
4. Use a Limits function.

The output of the staging program was a comprehensive XML file that was then processed using an XSL translator into the appropriate load statements for the production database. This intermediate XML step provided a wealth of information during development, and provides a convenient human- or machine-readable resource for the information in the database.

The Phase 5.3 data was invaluable in determining river and land segmentation, and plant uptake values. Data was available in comma-separated files in five-year increments from 1987 to 2002, and 1985 and 2005. BMP functions were derived from the same data set, but in a slightly less direct fashion. BMP usages are available as estimates of total acres per FIPS code. Efficiencies are available as the number of pounds of nutrients removed on a yearly basis. The average BMP usage per acre was calculated and used as the coverage for that FIPS code. Efficiency was calculated by dividing the pounds of nutrients removed by the BMP acres .

Land development functions were also based on the Phase 5.3 model data files over the same period. The Phase 5.3 model has a methodology for converting data from sources such as satellite and aerial photography into land usage estimates for every FIPS code in the watershed. Land segments are generally split

on county lines, though this rule was broken when significant physio- or oro-graphic differences exist within a county. Segments that aligned with the underlying county completely were given a prefix of A. When such an alignment was not possible, the FIPS code was prefixed with an A, B, or C to distinguish the segments (EPA, 2010#3). This segmentation fit well with the framework's land segment division principles.

The Phase 5.3 defines 26 separate land usages based on their methodology. However, many of these land usages are simply variations on a theme (EPA, 2010#3). For example, the Phase 5.3 data files define forest and harvested forest as separate land usages. As described above, the framework uses a single “forest” agent class to encompass both usages. The agent class has a member variable to describe the harvesting state of the area. Similar conglomerations were performed for nearly all Phase 5.3 land usages. All Phase 5.3 land usage types were accounted in the framework without loss of fidelity.

The USDA Economic Research Service was used to determine farm sizes in the prototype. The service provides important data of farm sizes throughout the states in the watershed (USDA, 2009). Farm size distributions in the prototype were identical to USDA statistics for number of farms per size category, average size, and median size. It was assumed that the statewide ratios would be consistent for portions of the state to account for two data inconsistencies:

1. Not all land segments are completely within the watershed.
2. The USDA data does not follow the Phase 5.3 model's land segmentation methodology.

Population data was derived from the US Census Bureau's population estimates for 1980 to 2009, including official Census numbers from 1980 and 1990. Regressions were fit using the 30 data points for each FIPS code.

Annual river segment flow rates were calculated from values provided from the Chesapeake Bay Program's website(CBPN, 2010)(CBPP, 2010). A linear flow multiplier function was used for every river segment. Baseline flow rates were calculated for each year as the ratio of that year's flow to the harmonic mean flow rate over all years.

The framework will be configured to monitor organic nitrogen and phosphorus, phosphates, nitrates, and ammonia. However, results reported here are computed as total nitrogen and phosphorus, and not broken out for each individual nutrient. All nitrogen EOF functions gave a constant baseline transfer rate of 60%. Phosphorus and phosphates used a constant transfer rate of 10%. EOF values for phosphorus and nitrogen are generally not available without site-specific study (FAPRI, 2007)(SERA, 2003).

Methodology

The Chesapeake Bay Watershed comprises a massive area with an enormous number of point and non-point sources of nutrients. The current scientific consensus is that nitrogen and phosphorus are decreasing the health of the Chesapeake Bay(EPA, 2010). However there is less consensus on how to manage the sources to improve the Bay health.

In an effort to gain insight on the dynamics of the watershed and showcase the flexibility of the framework, we performed model runs with all parameters at

their initial load values. We then varied river segment flow multipliers with constant EOF values, and EOF multipliers with constant river segment flow multiplier. The following functions were employed:

EOF	Nitrogen EOF	Phosphorus EOF
1	$f(x)=0.6$	$f(x)=0.1$
2	$f(x)=0.1x+0.5$	$f(x)=0.1x$
3	$f(x)=\begin{cases} x<0.7:0.2x+0.46 \\ 0.7\leq x\leq 1.3:0.6 \\ x>1.3:0.2x+0.34 \end{cases}$	$f(x)=\begin{cases} x<0.7:0.2x-0.04 \\ 0.7\leq x\leq 1.3:0.1 \\ x>1.3:0.2x-0.16 \end{cases}$
4	$f(x)=\begin{cases} Farm, CAFO:0.5 \\ Forest:0.2 \\ Other:0.9 \end{cases}$	$f(x)=\begin{cases} Farm, CAFO:0.08 \\ Forest:0.01 \\ Other:0.9 \end{cases}$

Flow	Nitrogen Effects	Phosphorus Effects
A	$f(x)=x$	$f(x)=x$
B	$f(x)=2x-0.75$	$f(x)=2x-0.75$
C	$f(x)=0.58*10.62^x$	$f(x)=0.58*10.62^x$

Population was also varied to gauge the effects of different population scenarios on the watershed. The population experiments were performed from 1990-2025 to forecast future changes to the watershed. Each population run was simulated with constant EOF and flow multiplier functions to eliminate those variables. The following scenarios were explored:

1. Zero population growth since 1990
2. Increased population growth by 5%
3. Uniformly distributed population

The model runs were performed using the UVA Fir cluster. The simulation period was 1990-2009. Results were then verified against published Phase 5.3

model results for the same period. All runs were started with the same random seed.

Prototype Results

The baseline framework values showed modest fidelity to the published Phase 5.3 model results for nitrogen, but did not closely mimic phosphorus totals. The best mean average percent error (MAPE) of the nitrogen loads (scenario A) was 7%, while the

best MAPE for

phosphorus loads

(scenario C) was 21%.

These results show

that a simple linear

function of flow was

most effective at

duplicating nitrogen

loads, while an

exponential function

worked best for

phosphorus loadings.

These results were

presaged by Phase 5.3

results, which show

strong correlation with *Figure 7: Flow Effects on Phosphorus*

the harmonic mean of water volume over the simulated period. Nitrogen loads

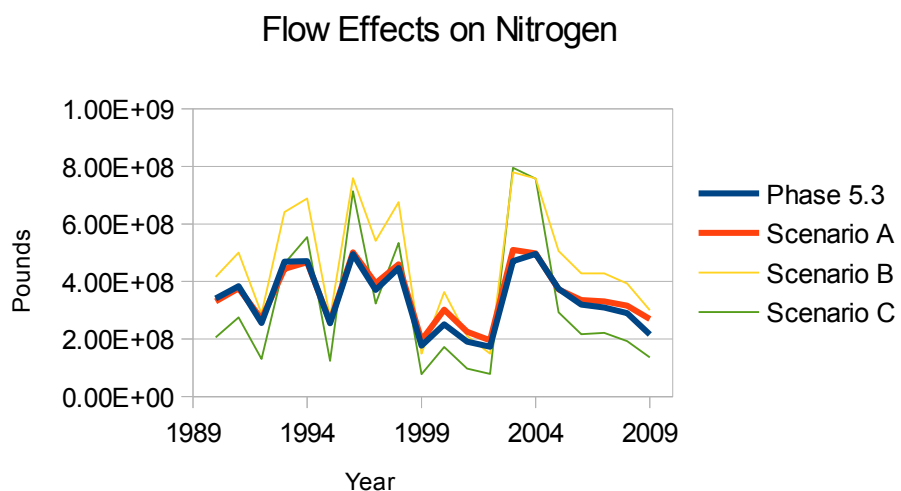


Figure 6: Flow Effects on Nitrogen

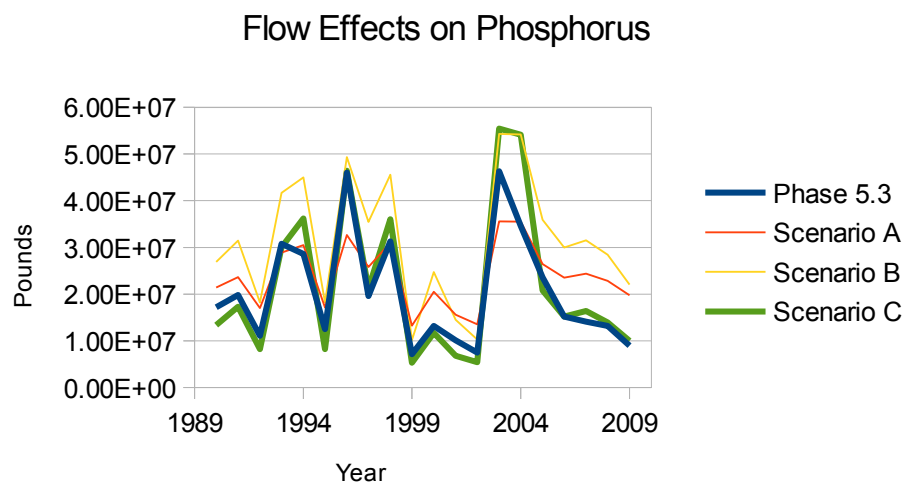


Figure 7: Flow Effects on Phosphorus

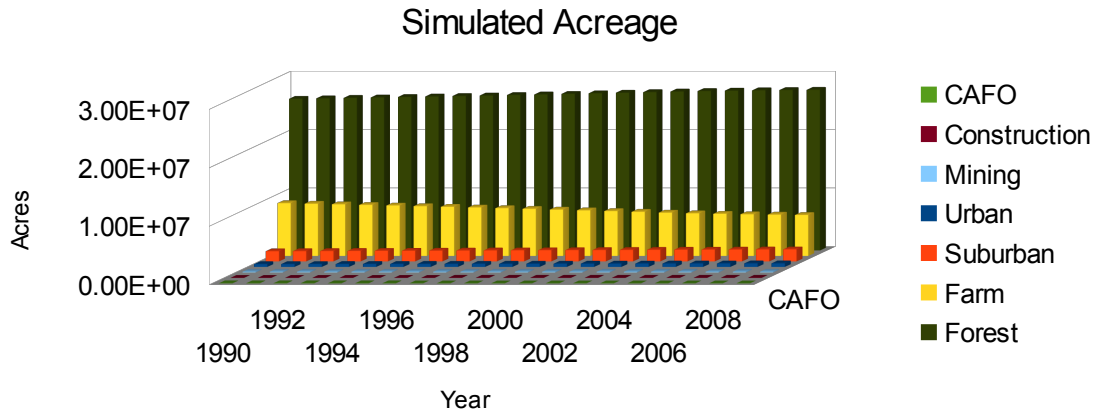


Figure 12: Simulated Acreage, Scenario 1

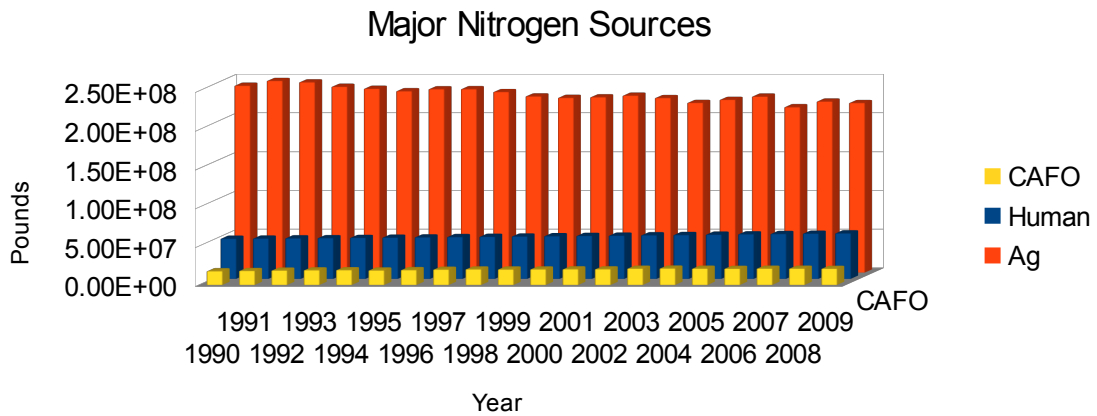


Figure 13: Major Nitrogen Sources, Scenario 1

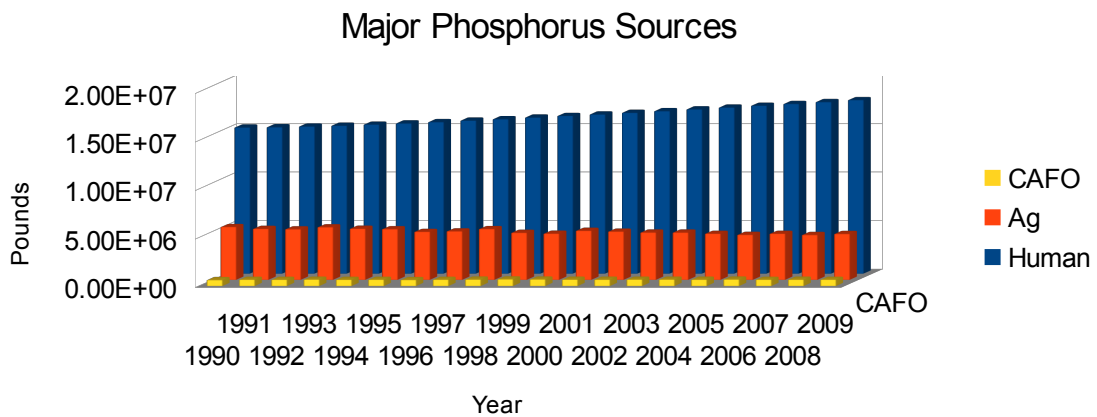


Figure 14: Major Phosphorus Sources, Scenario 1

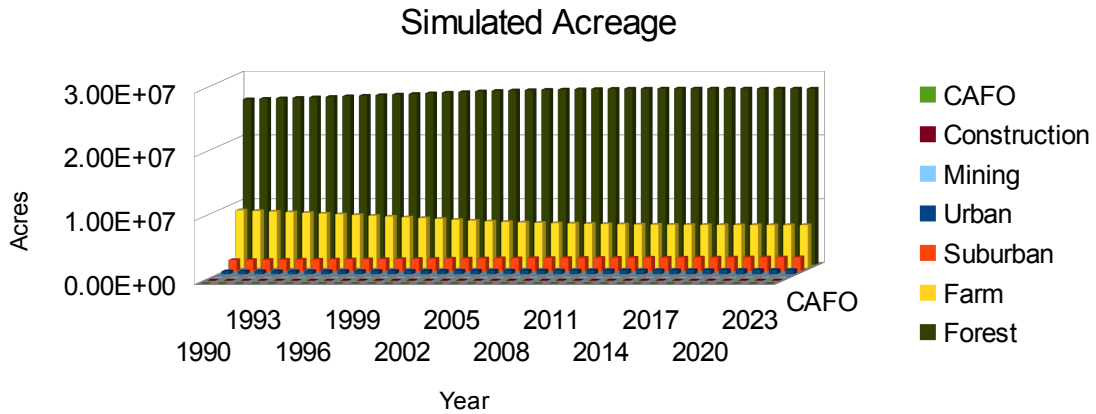


Figure 15: Simulated Acreage, Population Baseline

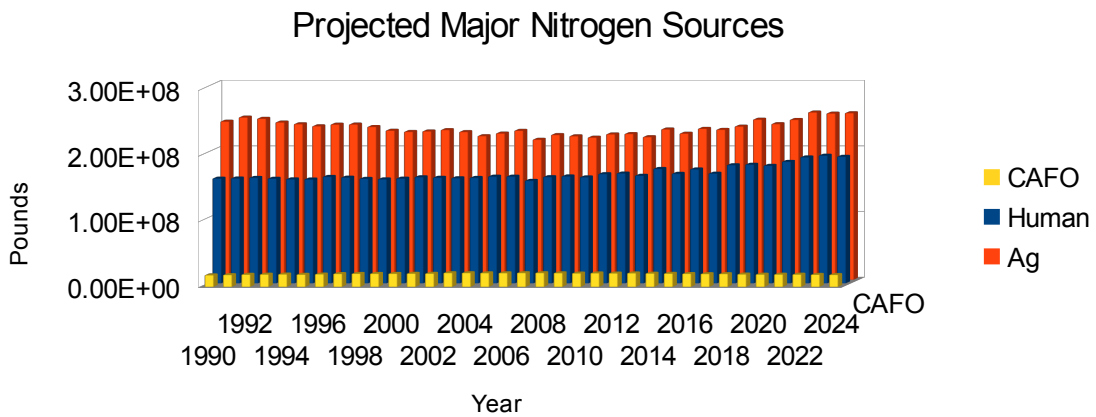


Figure 16: Major Nitrogen Sources, Population Baseline

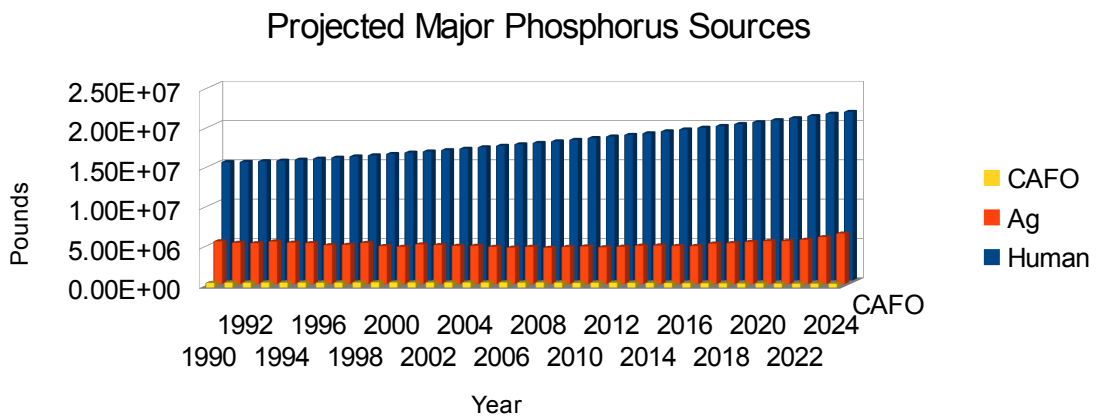


Figure 17: Major Phosphorus Sources, Population Baseline

have a linear fit with coefficient of determination $R^2=0.96$, and phosphorus loads have an exponential fit with coefficient of determination $R^2=0.95$. The framework supports unique flow multipliers per nutrient, so a single run can duplicate both linear and exponential scalings.

Interestingly, eliminating flow and EOF variability (scenario 1) in the framework revealed slightly decreasing nitrogen and slightly increasing phosphorus loads during the simulation. The nitrogen trajectory appears ephemeral in the face

of increased

population in the

watershed. Indeed,

the reductions in

farmland acreage and

the accompanying

reduction in nitrogen

load are offset by

increased population

and CAFO

contributions. This

trend is likely to

continue. However,

population numbers

are increasing

exponentially, while farmland appears on a roughly linear downward slope. In the

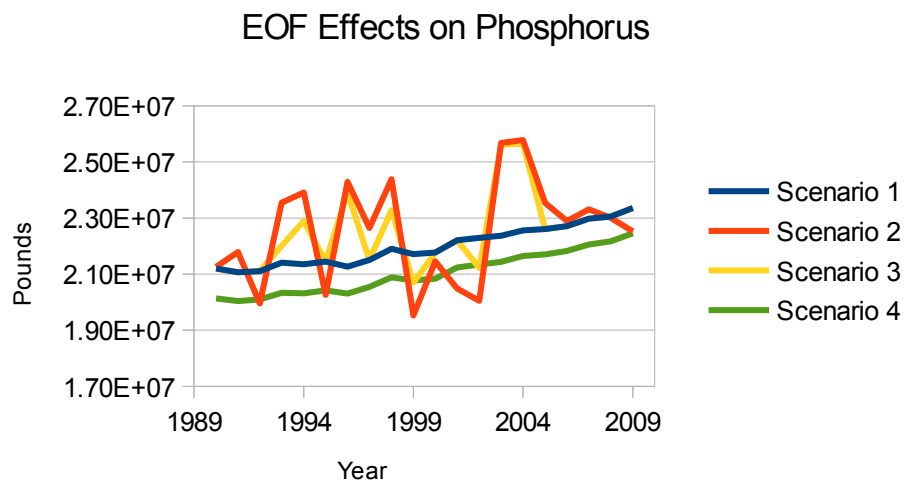


Figure 8: EOF Effects on Phosphorus

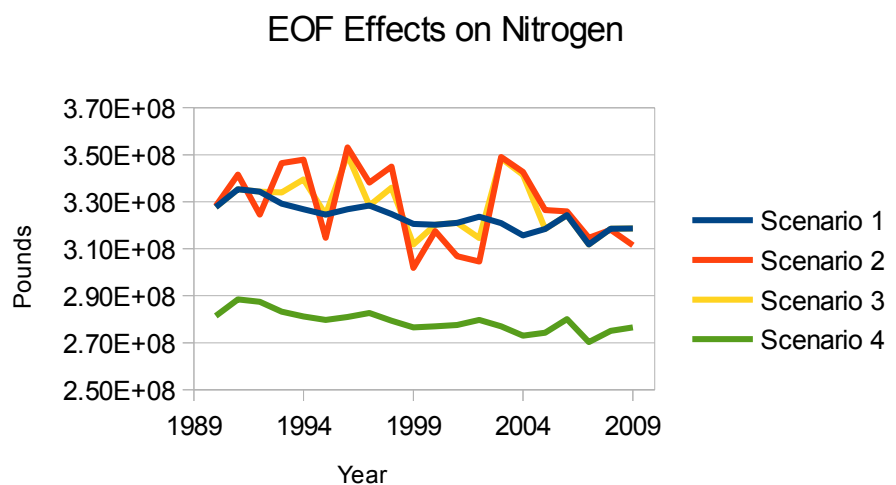


Figure 9: EOF Effects on Nitrogen

short term, the reductions from agriculture will continue to buffer the rising population numbers. Long-term projections are not expected to be promising, as detailed below. The phosphorus loadings follow the same pattern: decreasing load from agriculture is offset by exponentially increasing population contributions. In the case of phosphorus, however, agriculture's contribution is already overwhelmed by population's contribution.

Varying the other EOF multipliers had only a minor effect on the overall nutrient loadings. Scenario 2 was designed to explore a sliding EOF effect based on river segment flow.

Scenario 3 was a variation on a theme, but remained constant through one standard deviation of river flows. As expected,

the sliding scale scenario was more volatile than the constant values of scenario 1. Scenario 3 moderated the volatility of scenario 2. The per-land use EOF

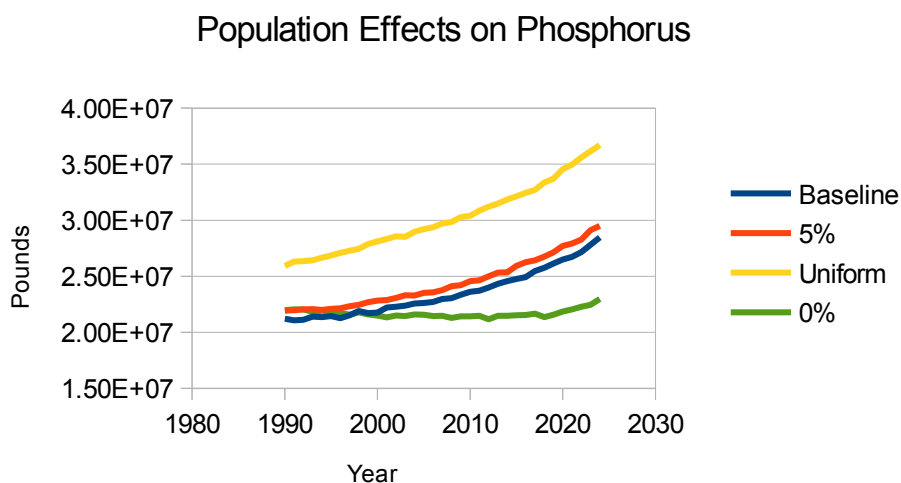


Figure 10: Population Effects on Phosphorus

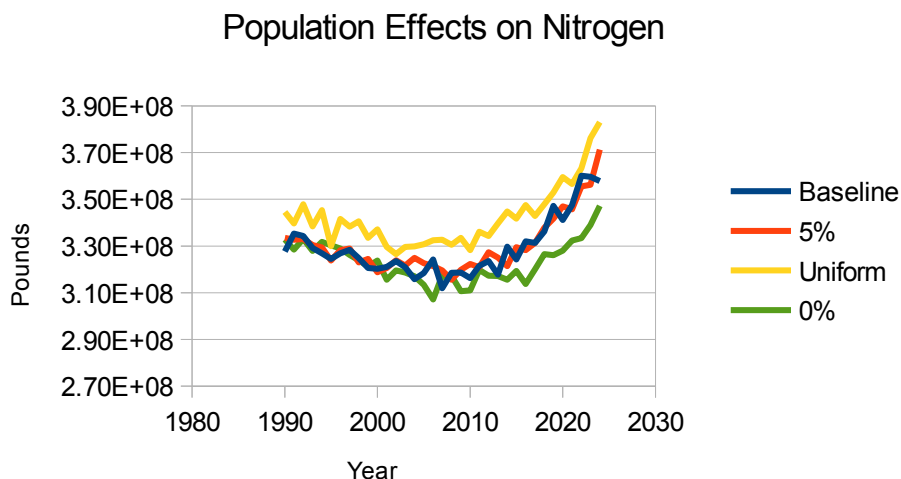


Figure 11: Population Effects on Nitrogen

(scenario 4) provides a possible proxy for measuring the effects of land usage changes. As land changes from one use to another, the nutrient loading profile changes, but so does its export risk. The hypothetical EOF values used in the simulation exhibit a 17% reduction in nitrogen contributions over the baseline numbers of scenario 1.

The population experiments forecast a troubled future for the Chesapeake Bay Watershed. While agricultural land usage is projected to continue its decline, the rate slows around 2015, and the load per acre increases for both nitrogen and phosphorus. The population contributions continue to increase. CAFO contributions peak in 2009 and then decline slightly over the rest of the simulation. Other nutrient sources remain insignificant. Interestingly, the uniform population scenario shows the greatest increase for both nutrients. This is likely a result of an increased rural population without access to municipal sewer systems found in suburban and urban areas.

The situation is most salient in the phosphorus loadings. Here, with population pressures already driving the nutrient loads, the levels rise exponentially for all but the zero population growth scenario. The uniform population scenario shows a 50% increase over the baseline simulation. The 5% increased population shows an expected 5% increase over the baseline. All non-zero population scenarios show roughly the same growth rates.

The nitrogen results are less dramatic. All scenarios exhibit increased loadings, reversing the trends visible in the 1990-2009 simulations. However, because population is not the main driver of nitrogen loads, the population increases are less pronounced within the totals. The zero population growth

scenario shows the smallest increases, while the uniform population scenario has the most. The baseline and 5% scenarios were indistinguishable in these simulations.

IBM World Community Grid

The prototype application of the framework was selected for inclusion on the IBM World Community Grid project. WCG provides roughly 210 years of CPU time per day to active research projects. There are currently nine active projects on the WCG (IBM, 2012). The WCG greatly expanded the computing power available to the prototype investigation. This framework will be launched on the WCG in April, 2012. Results have not yet been made available to researchers.

The additional computing power makes it possible to explore a much wider set of parameters and execute more than one run of each. Executing multiple runs of the same parameter set provides a baseline for detecting errors or otherwise anomalous results.

Data sets will be generated for the WCG project in three distinct sets. The first set will consist of exploring 18 distinct parameters, such as household size, or BMP usages. For each parameter, a high, low, and central value will be determined. Multiple runs will be performed for each high and low value, generating a response surface for that parameter. This set of experiments will contain 2^{18} unique sets of parameters.

The hyperplane developed in the first tranche of experiments will inform the design of the second, fractional factorial data set. This data set will more fully explore the interdependencies and relationships between the parameters by

running each parameter across a range of values against all other parameters in a range of values. Currently, we anticipate using a 3^{18-6} design, or approximately 5.3×10^5 sets of unique parameters. It will direct the design of the third data set.

The third data set will be a full factorial data set with a reduced set of parameters. Results from the fractional data set will determine which parameters are most likely to show the largest impact on the watershed. The WCG will provide the computational power to select 12 parameters for detailed examination. This will result in 3^{12} sets of unique parameters, minus some number of previously-computed sets from the second tranche. With the full factorial data set results in hand, we will be more likely to determine the optimum behaviors and policies to improve the health of the Chesapeake Bay Watershed.

Conclusion

Watersheds across the globe are under tremendous pressures from a number of sources: population growth, land development, excess nutrients, and overfishing, among others. Understanding how these pressures interact is of great importance to the vitality of the watershed. An agent-based model of the important factors contributing to the watershed health is a key step in understanding these interactions. At the same time, a working watershed model gives policy makers a tool in attempting to mitigate the deleterious effects.

There is a number of watershed modeling tools available to the watershed modeler or policy maker. However, a broader framework is needed to focus the work of many scientists, while freeing them of the need to "reinvent the wheel" for every watershed. This thesis describes an extensible agent-based model framework for modeling nutrient flows in a watershed.

Because of its proximity and readily available data, the Chesapeake Bay Watershed was used as the initial testbed for this framework. The system is database-driven, with easily modifiable parameters. The framework includes a generic concept of “nutrient,” and an extensible one of “function.” The model focuses on land-based non-point nutrient contributions, which are perhaps the most important and least understood nutrient contributions in a watershed. The model tracks nutrients from various land-based sources through the river system, and into the watershed.

Data acquisition remains a challenging aspect of watershed modeling. The framework's reliance on standard SQL eases data maintenance on the scientist, but does not alleviate the need to formulate functions describing the watershed. Relevant data is not always available or reliable, but it most directly affects the simulation results.

There is a tension between equation-based and agent-based modeling techniques, which is difficult to overcome. Equation-based models attempt to describe a system with a set of equations, while agent-based models attempt to describe the agents and behaviors that produce the system (Van Dyke Parunak, 1998). Merging the two techniques, as we have attempted with this framework, creates an equation-based system with an unaccountable variance in the results, or an agent-based system that is restricted to very narrow behaviors. Understanding and utilizing this tension remains an exciting area of further research.

Based on the results of the prototype application of the framework, it is clear that agents must be able to more easily interact with each other and other pieces of the framework. Without more complex interactions between agents, and the environment, there is little opportunity for the agents to alter their behavior. The possibility exists to make agents more aware of the overall health of the watershed, but this implies a more omniscient watershed modeler, and is likely to compromise the benefit of the agent-based model.

Bibliography

- US EPA(2002).*A framework for assessing and reporting on ecological condition*. Retrieved from <http://water.epa.gov/polwaste/nps/watershed/framework.cfm>.
- Booth DB & Jackson CR.(2001).*Land use and watersheds: human influence on hydrology and geomorphology in urban and forest areas*. Wignmosts MS & Burges SJ (Eds.). Washington, DC:American Geophysical Union.
- Ecorse Creek Inter-Municipality Commission(2001).*Ecorse creek watershed management plan*. Retrieved from http://www.ecorsecreek.com/wmp/ch7_050506.pdf.
- Alliance for the Great Lakes(2009).*Stresses and opportunities in illinois lake michigan watersheds*. Retrieved from <http://www.greatlakes.org/page.aspx?pid=881>.
- United States Environmental Protection Agency(2012).*Watershed assessment, tracking & environmental results*. Retrieved from http://iaspub.epa.gov/waters10/attains_nation_cy.control?p_report_type=T.
- Hallock D(2002).A water quality index for ecology's stream monitoring program(02-03-052)P.O. Box, 47600, Olympia, WA 98504:Washington State Department of Ecology
- Sánchez E, Colmenarejo MF, Vicente J, Rubio A, García MG, Travieso L & Borja R(2007).Use of the water quality index and dissolved oxygen deficit as simple indicators of watersheds pollution. *Ecological Indicators*,7,315-328.
- McClelland NI(1974).Water quality index application in the kansas river basin(EPA-907/9-74-001)Kansas City, MO:US Environmental Protection Agency
- Chesapeake Bay Program(2009).*History of the chesapeake bay program*. Retrieved from <http://www.chesapeakebay.net/historyofcbp.aspx?menuitem=14904>.
- Levin SA, Grenfell B, Hastings A & Perelson AS(2007).Mathematical and computational challenges in population biology and ecosystem science. *Science*,275,334-343.
- Aschmann SG, Anderson DP, Croft RJ & Cassell EA(1999).Using a watershed nutrient dynamics model, wend, to address watershed-scale. *Journal of Soil and Water Conservation*,54,630-635.
- USGS(1996).*Summary of hspf*. Retrieved from <http://water/usgs.gov/software/hspf.html>.
- United States Environmental Protection Agency(2007).*Basins 4.0-fact sheet*. Retrieved from <http://water.epa.gov/scitech/datait/models/basins/fs-basins4.cfm>.
- Chesapeake Bay Foundation(2008).*Bad water and the decline of blue crabs in the chesapeake bay*. Retrieved from <http://cbf.org/badwaters>.
- Anderson DM, Gilbert PM & Burkholder JM(2002).Harmful algal blooms and eutrophication: nutrient sources, composition, and consequences. *Estuaries*,25,704-726.
- Bowen JL & Valiela I(2001).The ecological effects of urbanization of coastal

- watersheds: historical increases in nitrogen loads and eutrophication of waquoit bay estuaries. *Canadian Journal of Fisheries and Aquatic Sciences*,58,1489-1500.
- Breitburg D(2002).Effects of hypoxia, and the balance between hypoxia and enrichment, on coas. *Estuaries*,5,767-781.
- Smith RA & Alexander RB(2000).Sources of nutrients in the nation's watersheds(unnumbered)Reston, VA:US Geological Survey
- Burkholder J, Libra B, Weyer P, Heathcote S, Kolpin D, Thorne PS & Wichman M(2007).Impacts of waste from concentrated animal feeding operations on water quality. *Environmental Health Perspectives*,115,308-312.
- United Nations Population Fund(2007).*State of world population 2007*. Retrieved from <http://www.unfpa.org/swp/2007/english/introduction.html>.
- US EPA(2001).Our built and natural environments(EPA 231-R-01-002)Washington, DC:US Environmental Protection Area
- Tang Z, Engel BA, Pijanowski BC & Lim K(2005).Forecasting land use change and its environmental impact at a watershed scale. *Journal of Environmental Management*,76,35-45.
- Schueler T(1995).Environmental land planning series: site planning for urban stream protection(95708)Washington, DC:Center for Watershed Protection
- La Freenierre J(unpublished) La Freenierre J. The relationship between land change and water resources vulnerability: a review of existing literature.
- Patric JH & Reinhart KG(1971).Hydrologic effects of deforesting two mountain watersheds in west virginia. *Water Resources Research*,7,1182-1188.
- Owens DW, Jopke P, Hall DW, Balousey J & Roa A(2000).Soil erosion from two small construction sites, dane county, wisconsin(FS-109-00)Middleton, WI:United States Geological Survey
- Rawat JS & Rawat MS(1994).Accelerated erosion and denudation in the nana kosi watershed, central himalaya, india. part i: sediment load. *Mountain Research and Development*,14,25-38.
- Henley WF, Patterson MA, Neves RJ & Lemly AD(2000).Effects of sedimentation and turbidity on lotic food webs: a concise review for natural resource managers. *Reviews in Fisheries Science*,8,125-139.
- Mallin MA, Williams KE, Esham EC & Lowe RP(2000).Effects of human development on bacteriological water quality in coastal watersheds. *Ecological Applications*,10,1047-1056.
- Steffy LY & Kilham SS(2004).Elevated $\delta^{15}\text{N}$ in stream biota in areas with septic tank systems in an urban watershed. *Ecological Applications*,14,637-641.
- Robinson D(2001).Delta(^{15}N) as an integrator of the nitrogen cycle. *Trends in Ecological Evolution* **16**: 153-162.
- Fuller BT, Fuller JL, Sage NE, Harris DA, O'Connel TC & Hedges REM(2005).Nitrogen balance and delta ^{15}N : why you're not what you eat during pregnancy. *Rapid Communincations in Mass Spectrometry*,19,2497-2506.
- Healthy Harbor(2001).*Healthy harbor plan*. Retrieved from http://www.healthyharborbaltimore.org/uploads/file/03_Chapter_3_Sewage.pdf.
- Jha S & Bawa KS(2006).Population growth, human development, and deforestation in biodiversity hotspots. *Conservation Biology*,20,906-912.

- Wood CH & Skole DL.(1998).*Linking satellite, census, and survey data to study deforestation in the brazilian amazon*. D. Liverman (Ed.). Washington, DC:National Academies Press.
- Root TL, Price JT, Hall KR, Schneider SH, Rosenzweig C & Pounds JA(2003).. *Nature*,421,57-60.
- Trenberth K(2005).Uncertainty in hurricanes and global warming. *Science*,308,1753-1754.
- Adams RM, Rosenzweig C, Peart RM, Ritchie JT, McCarl BA, Glycer JD, Curry RB, Jones JW, Boote KJ & Allen LH(1990).Global climate change and us agriculture. *Nature*,345,219 - 224.
- US Environmental Protection Agency(2010).Chesapeake bay phase 5 community watershed model in preparation(EPA XXX-X-XX-010)Annapolis, MD:Chesapeake Bay Program Office
- US Environmental Protection Agency(2010).Chesapeake bay phase 5 community watershed model in preparation(EPA XXX-X-XX-010)Annapolis, MD:Chesapeake Bay Program Office
- US Environmental Protection Agency(2010).Chesapeake bay phase 5 community watershed model in preparation(EPA XXX-X-XX-010)Annapolis, MD:Chesapeake Bay Program Office
- US Environmental Protection Agency(2010).Chesapeake bay phase 5 community watershed model in preparation(EPA XXX-X-XX-010)Annapolis, MD:Chesapeake Bay Program Office
- Band L, Dillaha T, Diffy C, Reckhow K & Welty C(2008).Chesapeake bay watershed model phase v review(STAC-08-003):Chesapeake Bay Foundation
- Scientific Software Group(2012).*Wms-watershed modeling system*. Retrieved from http://www.scisoftware.com/environmental_software/product_info.php?products_id=120.
- Ottino JM(2004).Complex systems. *AIChE Journal* **49**: 292-299.
- Berry BJL, Kiel DL & Elliott E(2002).Adaptive agents, intelligence, and emergent human organization: capturing complexity through agent-based modeling. *Proceedings of the National Academy of Social Sciences of the United States of America* **99**: 7187-7188.
- Van Dyke Parunak H, Savit R & Riolo RL(1998).Agent-based modeling vs. equation-based modeling: a case study and users' guide. *Proceedings of Multi-agent Systems and Agent-Based Simulations (MABS)*.(LNAI 1434).10-25.
- Jennings NR(2001).An agent-based approach for building complex software systems. *Communications of the ACM* **44**: 35-41.
- Codd EF(1970).A relational model of data for large shared data banks. *Communications of the ACM* **13**: 377-387.
- Wikipedia(2012).*Object database*. Retrieved from http://en.wikipedia.org/wiki/Object_database.
- Wikipedia(2012).*Nosql*. Retrieved from <http://en.wikipedia.org/wiki/NoSQL>.
- Date CJ & Darwin H.(1997).*A guide to the sql standard : a user's guide to the standard database language sql*. . Indianapolis, IN:Addison-Wesley Professional.
- Hrudley SE, Gosselin P, Naeth MA, Plourde A, Therrien R, Van Der Kraak G & Xu Z(2010).The royal society of canada expert panel: environmental and health

- impacts of canada's oil sands industry(December, 2005):Royal Society of Canada
- Geosyntec Consultants(2008) Geosyntec Consultants. Bmp effectiveness assessment for highway runoff in western washington.
- Park SW, Mostaghimi S, Cooke RA & McClellan PW(1994).Bmp impacts on watershed runoff, sediment, and nutrient yields. *Journal of the American Water Resources Association* **30**: 1011-1023.
- Ekman M, Warg F & Nilsson J(2004).An in-depth look at computer performance growth(2004-9)Goteburg, Germany:Chalmers University of Technology
- Doran J.(2006).*Agent based computational modelling: applications in demography, social, economic and environmental sciences*. Billari FC, Fent T, Prskawetz A & Scheffran J (Eds.). Heidelberg and New York:Phsyica-Verlag.
- Milesi C & Running SW(2005).Mapping and modeling the biogeochemical cycling of turf grasses in the united states. *Environmental Management* **36**: 426-438.
- State Water Control Board(2012).*Vpdes watershed general permit for nutrient discharges to the chesapeake bay*. Retrieved from <http://www.deq.virginia.gov/export/sites/default/vpdes/pdf/9VAC25-820-NutrientDischargesGP2012.pdf>.
- Carley K, Fridsma D, Yahja A, Altman N, Chen L & Kaminsky B(2006).Biowar: scalable agent-based model of bioattacks. *IEEE Transactions on Systems, Man, and Cybernetics, Part A*,36,252-265.
- Schonfisch B & de Roos A(1999).Synchronous and asynchronous updating in cellular automata. *BioSystems* : 123-143.
- Welch L & Ekwaro-Osire S(2010).Fairness in agent based simulation frameworks. *J. Comput. Inf. Sci. Eng.*,10,21-28.
- Brezonik PL & Stadelmann TH(2002).Analysis and predictive models of stormwater runoff volumes, loads, and pollutant concentrations from watersheds in the twin cities metropolitan area, minnesota, usa. *Water Research* **36**: 1743-1757.
- Chesapeake Bay Foundation(2010).*Mission and vision*. Retrieved from <http://www.cbf.org/page.aspx?pid=387>.
- United States Department of Agriculture(2009).*Ers/usda data sets*. Retrieved from <http://www.ers.usda.gov/Data/>.
- Chesapeake Bay Program(2010).*Track the progress*. Retrieved from http://www.chesapeakebay.net/indicators/indicator/nitrogen_loads_and_river_flow_to_the_bay1.
- Chesapeake Bay Program(2010).*Track the progress*. Retrieved from http://www.chesapeakebay.net/indicators/indicator/phosphorus_loads_and_river_flow_to_the_bay.
- Food and Agricultural Policy Research Institute.(2007).*Estimating water quality, air quality, and soil carbon benefits of the conservation reserve program*. . University of Missouri - Columbia:Food and Agricultural Policy Research Institute.
- Radcliffe D & Nelson N(2003).*Predicting phosphorus losses*. Retrieved from www.sera17.ext.vt.edu/Documents/Predicting%20P%20Losses.pdf.
- IBM(2012).*Global statistics*. Retrieved from <http://www.worldcommunitygrid.org/stat/viewGlobal.do>.

