

SIMULATION

<http://sim.sagepub.com/>

Watershed modeling using large-scale distributed computing in Condor and the Soil and Water Assessment Tool model

Margaret W Gitau, Li-Chi Chiang, Mohamed Sayeed and Indrajeet Chaubey

SIMULATION published online 16 May 2011

DOI: 10.1177/0037549711402524

The online version of this article can be found at:

<http://sim.sagepub.com/content/early/2011/05/13/0037549711402524>

Published by:



<http://www.sagepublications.com>

On behalf of:



Society for Modeling and Simulation International (SCS)

Additional services and information for *SIMULATION* can be found at:

Email Alerts: <http://sim.sagepub.com/cgi/alerts>

Subscriptions: <http://sim.sagepub.com/subscriptions>

Reprints: <http://www.sagepub.com/journalsReprints.nav>

Permissions: <http://www.sagepub.com/journalsPermissions.nav>



Watershed modeling using large-scale distributed computing in Condor and the Soil and Water Assessment Tool model

Margaret W Gitau¹, Li-Chi Chiang², Mohamed Sayeed³ and Indrajeet Chaubey⁴

Abstract

Models are increasingly being used to quantify the effects of best management practices (BMPs) on water quality. While these models offer the ability to study multiple BMP scenarios, and to analyze impacts of various management decisions on watershed response, associated analyses can be very computationally intensive due to a large number of runs needed to fully capture the various uncertainties in the model outputs. There is, thus, the need to develop suitable and efficient techniques to handle such comprehensive model evaluations. We demonstrate a novel approach to accomplish a large number of model runs with Condor, a distributed high-throughput computing framework for model runs with the Soil and Water Assessment Tool (SWAT) model. This application required more than 43,000 runs of the SWAT model to evaluate the impacts of 172 different watershed management decisions combined with weather uncertainty on water quality. The SWAT model was run in the Condor environment implemented on the TeraGrid. This framework significantly reduced the model run time from 2.5 years to 18 days and enabled us to perform comprehensive BMP analyses that may not have been possible with traditional model runs on a few desktop computers. The Condor system can be used effectively to make Monte Carlo analyses of complex watershed models requiring a large number of computational cycles.

Keywords

best management practices, Condor, Conservation Effectiveness Assessment Program, Lincoln Lake, Soil and Water Assessment Tool Model, TeraGrid

1. Introduction

The interactions among watershed geophysical attributes (e.g. land use, soils, topography), climate, land management, and their impacts on hydrologic and water quality response at various spatial and temporal scales are highly complex and non-linear. Given these complexities, watershed models are often used to quantify these interactions. Similarly, watershed models are increasingly used to evaluate watershed response under various climate and management scenarios, and to make watershed management recommendations for improving water quality.¹ For example, many studies have used the modeling results to evaluate the effectiveness of best management practices (BMPs) and to determine the optimum BMPs, which can greatly improve water quality in watersheds.^{2–5}

The recent advances in computational resources, such as the availability of fast computers with large

¹Biological and Agricultural Systems Engineering, and Center for Water and Air Quality, Florida A&M University, USA.

²Department of Agricultural and Biological Engineering, Purdue University, USA.

³Computing Research Institute, Rosen Center for Advanced Computing, USA.

⁴Department of Agricultural and Biological Engineering, Department of Earth and Atmospheric Sciences, and Division of Environmental and Ecological Engineering, Purdue University, USA.

Corresponding author:

I Chaubey, Department of Agricultural and Biological Engineering, Department of Earth and Atmospheric Sciences, and Division of Environmental and Ecological Engineering, Purdue University, 225 South University Street, West Lafayette, IN 47907, USA
Email: ichaubey@purdue.edu

memory, have enabled greater representation of watershed heterogeneity in model development, and simulation of a greater number of management scenarios than was possible even a decade ago. However, there are still a number of computational challenges in applying complex watershed models to evaluate the impact of various watershed management decisions on watershed response functions. These challenges include: (1) inverse modeling to effectively identify model parameters through a model calibration process, so that watershed response predictions for future conditions can be made; and (2) quantifying model output uncertainty due to various input data and model parameters. In the inverse modeling, measured data are used to compare the model outputs and to adjust the model parameters until a satisfactory match between the measured and modeled data are obtained. A number of inverse modeling approaches have been developed for complex watershed models. These approaches are computationally very intensive and require a larger number of model runs. Model calibration to identify parameter values requires sensitivity analyses and parameter estimation to match model predictions with observed watershed response data using predefined objective functions. Although there are some parameters that have been generally found sensitive, there is some degree of site specificity related to base conditions used for various parameters,⁶ thus the need for site-specific sensitivity analyses.

Sensitivity analyses and autocalibration of the Soil and Water Assessment Tool (SWAT) model⁷ are reported to range from a few days to more than a month,⁸ primarily due to the very large number of model runs needed for parameter identification. Zhang et al.⁹ needed 10,000 model runs to estimate SWAT parameters using single- and multi-objective optimization methods. Similar computational requirements have been reported by Bekele and Nicklow¹⁰ for the SWAT model. Likewise, various researchers have documented the need for uncertainty quantification in model predictions from various sources of error, such as input variability, model algorithms, model calibration data, parameter variability, etc.^{11,12} Uncertainty analyses may require a large number of model runs and can be computationally expensive considering the size of the watershed, model representation, and period of simulation.¹³

Recently, there has been considerable interest in quantifying the impact of various BMPs in improving water quality using watershed models. The United States Department of Agriculture (USDA) has funded several Conservation Effectiveness Assessment Program (CEAP) studies to evaluate

how various BMPs interact in improving water quality at various spatial and temporal scales (<http://www.nrcs.usda.gov/technical/nri/ceap/index.html>). Many of these studies involve running watershed models, such as the SWAT to evaluate the interactions of watershed geophysical characteristics, climate, and BMPs on hydrologic/water quality response. The SWAT model has more than 200 parameters that describe various land and water phases of hydrologic cycles, and watershed response outputs. A large number of model parameters necessitate a large number of model runs for sensitivity analysis, parameter estimation, or for Monte Carlo assessment of output uncertainty. The run time for a single model run depends on the size of the watershed, watershed discretization, and length of the simulation period. Typically, it can range from a few minutes to more than 1 hour for an eight-digit hydrologic unit code (HUC) watershed. A sequential model run for parameter estimation, model calibration, or watershed response predictions can thus take a relatively long time, ranging from a few days to months.⁸ The application of a calibrated watershed model to evaluate the effectiveness of various BMPs in improving water quality can also be computationally cumbersome. For example, Maringanti et al.¹⁴ have reported that optimizing BMPs in agricultural watersheds using the SWAT model may require a very large number of model runs (10^{300} runs for an agricultural watershed having 500 farms and four BMPs possible for each farm). This computational burden is one of the primary reasons why the SWAT model has not been used to optimize BMPs in large agricultural watersheds.¹⁴

In this paper, we present the use of a high-throughput computing framework called Condor,^{15,16} a public domain high-throughput computing software system, to accomplish the large number of SWAT model runs for output uncertainty analyses, and BMP effectiveness assessment. Condor utilizes unused idle cycles of desktop machines or clusters,^{15,17} thus allowing a large number of sequential and parallel jobs to be executed. This application is demonstrated through a case study of evaluating the impact of 172 different watershed management decisions on watershed response under uncertain weather conditions. This application required more than 43,000 runs of the SWAT model. This novel application allowed us to efficiently perform a large number of SWAT runs in a complex watershed. It should be noted that the SWAT model results are provided here primarily as a case study to demonstrate the utility of the Condor system in making a large number of complex watershed model runs. Detailed SWAT results are reported in Chaubey et al.¹⁸ and Chiang et al.¹⁹

2. Material and methods

2.1. Study watershed description

This study was conducted in the Lincoln Lake Watershed, a 32 km² watershed located in the Ozark Highlands of the northwest Arkansas, USA (Figure 1). Moores Creek and Beatty Branch are two major tributaries that flow into Lincoln Lake (Figure 1). Dominant land uses in the watershed, based on 2004 Landsat data, include forest (39% of the total watershed area), pastures (36% of the watershed area), and urban (12% of the watershed area). Non-point source (NPS) transport of nutrients, sediment, and pathogens from agricultural activities is a major concern in this area.^{20,21} The rolling hills in this region are home to numerous poultry farms and pastures that produce abundant forage for numerous beef and dairy cattle. The predominant use of animal manure in the area has been as a

fertilizer for perennial forage crops. There is growing concern that excess land applications of animal manure can lead to surface and ground water pollution due to increased runoff losses of nutrients such as nitrogen (N) and phosphorus (P), sediment, and pathogens (e.g. Edwards et al.²²). Increasingly, watersheds are unable to utilize the high levels of fertilizers and animal manure applied to them. The result is increases in noxious, oxygen-consuming and sometimes toxic algal blooms, deteriorations of fisheries, and general degradation of water quality.²³

This watershed is one of the 13 CEAP watersheds funded by the USDA Cooperative State Research, Education, and Extension Service (USDA-CSREES) to evaluate the effectiveness of various BMPs in improving the water quality and to evaluate how BMPs interact at various spatial and temporal scales to affect water quality. To minimize NPS pollution and to improve water quality, several management

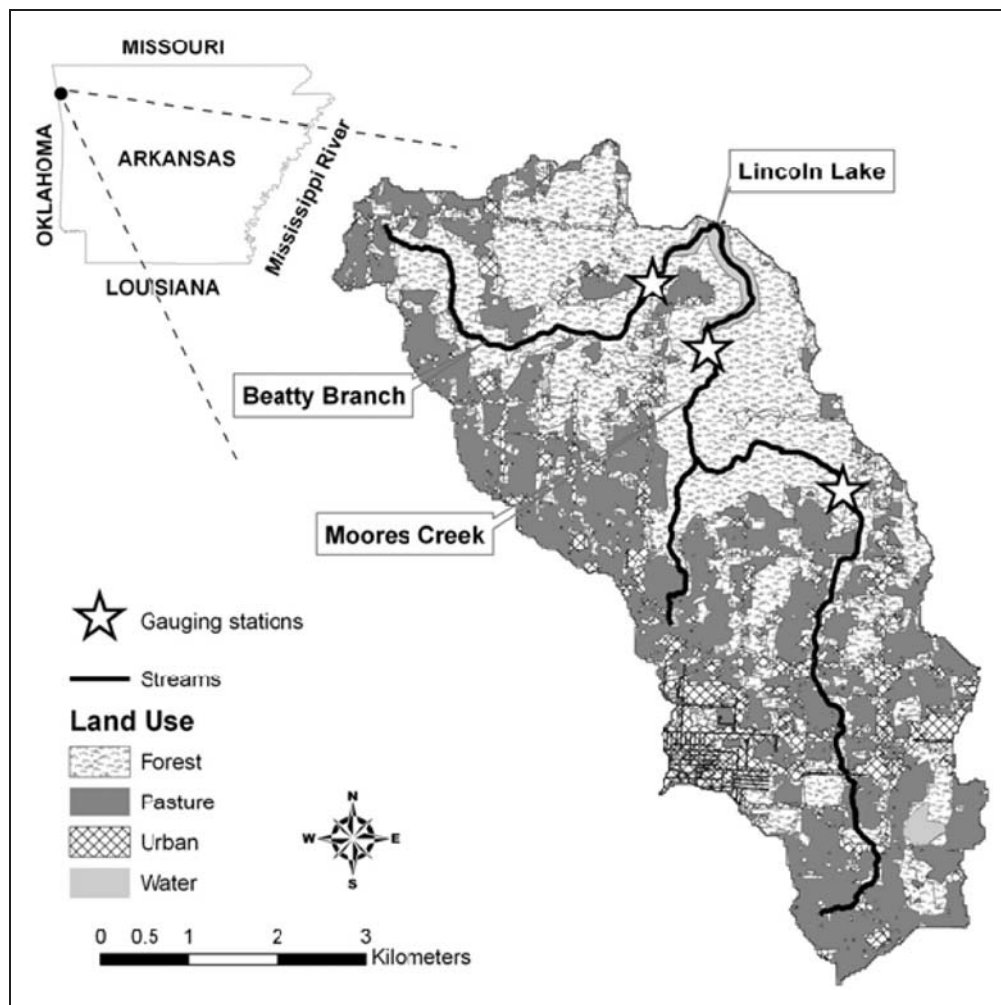


Figure 1. Location of the Lincoln Lake watershed, stream gauging stations, and land use.

practices have been adopted in the watershed over the past 15 years, including reduced poultry litter and commercial fertilizer application rates, application timing and chemical amendment to poultry litter, improved grazing and pasture management, and edge-of-field and riparian buffer zones. Discharge and water quality measurements have been done in the watershed to quantify concentrations and loads of various water quality parameters, including sediment and various forms of nitrogen (N) and phosphorus (P).

2.2. SWAT model description

The methodology is demonstrated using an example application of the SWAT model to evaluate performance of various management practices and stochasticity in future weather conditions in improving water quality. The SWAT model is a physically based, continuous simulation, daily time-step, distributed-parameter river basin, or watershed-scale hydrological model.^{7,24} It was designed to predict the impacts of land use and land management practices on water, sediment and agricultural chemical yields in complex watersheds over long periods of time. Detailed descriptions of this model have been presented previously by a number of authors, for example Arabi et al.,² Migliaccio and Chaubey,¹³ White and Chaubey,²⁵ Gitau et al.,^{4,26} and Gassman et al.,²⁷ among others. The model has the ability to provide watershed response output at various temporal (daily, monthly, and annual) and spatial (Hydrologic Response Unit or HRU, basin and sub-basin, stream reach) scales. The HRUs are the basic units upon which model computations are performed. The SWAT allows flexibility in subdividing sub-basins into unique HRUs based on user-specified land use and soil distribution thresholds. These thresholds determine the degree of lumping within the model and can both be set to zero to preclude lumping.

Of interest to this study were three key output files: the standard output file (output.std); the HRU level output file (output.hru); and the output at designated watershed outlets (out.out). Making such evaluations requires a large number of simulations of watershed models, thus necessitating the need for the development of this approach. For example, this application of the SWAT model to evaluate the performance of 172 different management practices under 250 simulated weather realizations required more than 43,000 runs of the SWAT model. The details about the model development, the Condor system, and execution in Condor are provided in the subsequent sections. Even though this application is demonstrated using the SWAT model, we, however, believe that the system so developed is applicable to other watershed models also,

provided that those models can be run on platforms supporting Condor.

The SWAT model was selected in part because of the flexibility it offers in discretization of the watershed and its ability to provide watershed response output at various temporal and spatial scales. Various automated geospatial tools and methods have been developed to prepare model input files^{28,29} and to calibrate model parameters.^{9,30} The SWAT model has been successfully applied to quantify the water quality impacts of various management decisions at various spatial scales ranging from fields^{26,31} to large watersheds.^{32,33} More than 350 peer-reviewed journal articles have been published demonstrating the capability of the SWAT model to evaluate watershed response at spatial scales ranging from fields to large watersheds and at temporal scales ranging from daily to several decades.²⁷ The SWAT model is currently being applied in many of the USDA CEAP watersheds to evaluate the watershed response under various BMP implementation conditions.

2.3. SWAT input development

For this study, 30-m United States Geological Survey (USGS) Digital Elevation Model (DEM) data were used for watershed delineation. Land use data with 30-m spatial resolution representing 2004 land use/cover in the watershed were obtained from the Center for Advanced Spatial Technology (CAST), University of Arkansas. Soil Survey Geographic (SSURGO, 1:12,000–1:63,360) level soils data were obtained from the Natural Resources Conservation Service (NRCS) Soils Data Mart.³⁴ In order to preclude lumping of different land use and soil categories, and to give the HRUs a spatial definition, HRUs were defined using the 0/0% land use and soil definition threshold.⁴ Giving the HRUs a spatial definition was instrumental in defining the baseline scenario, as it allowed more precise representation of current practice.

The simulation period of interest for this study was 2004–2028. The SWAT model was run from 2001 to 2028 with the initial three years of the simulation used as the model warm up period. Model simulations for future conditions require generation of weather data (e.g. precipitation, temperature, solar radiation, etc.). In order to incorporate the impact of stochasticity in weather on BMP performance, 250 different weather realizations were generated from 2001–2028 using WXGEN,³⁵ a weather generator software. WXGEN is also used by the SWAT model to generate weather data for future conditions or to fill the missing weather data for historical data. Within the SWAT, the generator uses a random seed to generate weather data at each run. In order to ensure that the same weather realization was used for all scenarios at each instant,

and also to improve computation efficiency, we obtained WXGEN as a stand-alone application and generated all weather data prior to performing the SWAT runs. Measured historical data (1990–2002) from the watershed were used as an input to the WXGEN model. The generated weather data included daily precipitation, solar radiation, wind speed, and minimum and maximum air temperatures.

This study considered 171 different BMP scenarios (Table 1). These 171 BMP scenarios included three general pasture management categories: (1) grazing and pasture management; (2) riparian and buffer zones; and (3) poultry litter and commercial fertilizer application practices (Table 1). These scenarios were selected based on numerous discussions with county extension agents and farmers and represent management options that can be adopted to manage pasture lands in this and other regional watersheds. Grazing and pasture

management comprised three levels: no grazing, optimum grazing, and over grazing. Based on information on typical optimal grazing management, we assumed that the cattle grazed through the entire watershed in 30 days and stayed for approximately 4–6 days in each pasture HRU.³⁶ The overgrazing application started on 30 September and lasted for 213 days until 30 April of each year. Three different widths of riparian/buffer zones were evaluated in this study (0, 15, 30 m). We assumed that the buffer strips were located at the end of pasture fields and received runoff from pasture areas. With regard to poultry litter and commercial fertilizer applications, three factors were considered: poultry litter application rates, litter characteristics (alum-amended litter and non-amended litter), and application timing (spring, summer, and fall). Litter application rates used in this study were 1, 1.5 and 2 tons/acre for spring (applied on 30 April) and summer (31 August)

Table 1. List of best management practice scenarios analyses in the Soil and Water Assessment Tool model

	Bffer width								
	0 m			30 m			15 m		
	Grazing and pasture management								
Manure application (tons/acre)	NG	OG	OVG	NG	OG	OVG	NG	OG	OVG
No application	1 ^a	20	39	58	77	96	115	134	153
Spring app									
1 A	2	21	40	59	78	97	116	135	154
1.5 A	3	22	41	60	79	98	117	136	155
2 A	4	23	42	61	80	99	118	137	156
1 NA	5	24	43	62	81	100	119	138	157
1.5 NA	6	25	44	63	82	101	120	139	158
2 NA	7	26	45	64	83	102	121	140	159
Summer									
1 A	8	27	46	65	84	103	122	141	160
1.5 A	9	28	47	66	85	104	123	142	161
2 A	10	29	48	67	86	105	124	143	162
1 NA	11	30	49	68	87	106	125	144	163
1.5 NA	12	31	50	69	88	107	126	145	164
2 NA	13	32	51	70	89	108	127	146	165
Fall									
2 A	14	33	52	71	90	109	128	147	166
2.5 A	15	34	53	72	91	110	129	148	167
3 A	16	35	54	73	92	111	130	149	168
2 NA	17	36	55	74	93	112	131	150	169
2.5 NA	18	37	56	75	94	113	132	151	170
3 NA	19	38	57	76	95	114	133	152	171

^aScenario numbers.

Grazing management (NG: no grazing, OG: optimum grazing, OVG: over grazing), buffer zone management (0, 15, 30 m) and nutrient management (spring, summer, and fall applications), where the number indicates the fertilizer amount in ton/acre unit with non-alum (NA) or alum (A) amended poultry litter.

applications, and 2, 2.5 and 3 tons/acre for fall applications (15 October), consistent with current nutrient management practice (NMP) recommendations in the watershed. More details about how the SWAT model simulates those BMPs can be found in Chaubey et al.¹⁸

In addition, a baseline scenario (scenario number 172) was built based on historical BMP data developed for the watershed⁸ and 2004 land use, land management, and soils data. The historical database comprises both spatial (parcel/field) and tabular data and covers BMPs implemented in the watershed from 1992 to 2006. In order to build the scenario, a spatial layer of the HRUs was first developed by overlaying the land use, soils, and sub-basin data as described by Gitau et al.⁴ and Gitau.³⁷ The BMP data was then overlaid with the HRU layer to determine which BMPs pertained to which HRUs. The BMPs were then input into their respective HRUs within the SWAT model. Where more than one HRU occurred within a field, the same data were used for the HRUs. Where more than one field with different management occurred within a SWAT-defined HRU, area-weighted values (for example of manure application rates) were used for the HRU.

2.4. Description of Condor – a high-throughput computing system

Condor is a free public domain software system for high-throughput computing. It allows harvesting of idle cycles of desktop machines or clusters.^{15–17,38} The system operates based on the notion that computers spend as much time idling as they do in use. Considering a university setting, for example, idle periods comprise the times when students are on break and, on a day-to-day basis, when students and other university personnel break off for lunch, or for the evenings and weekends. Condor seeks out such idle processors (CPUs) at run time and distributes model runs to these processors (Figure 2). The system gives priority to the owner such that any front-end activity on the executing computer (e.g. movement of the mouse) will result in job termination or migration. The Condor system can be configured to handle applications that are time intensive and involve large volumes of runs. The Condor system (head nodes) keeps track of the resources in the Condor pool, such as individual personal computers (PCs), lab PCs, and large clusters.

Condor provides two primary modes (or universes) to run jobs: the Standard and the Vanilla universe. The Standard universe provides a variety of advantages over the Vanilla universe, including checkpointing. This checkpoint mechanism allows process migration (handled internally in Condor), for example when the executing CPU is no longer available to the Condor pool and the running process has to stop, the Condor scheduler will allocate the job to a new machine.

This universe, however, requires compiling and re-linking the executable with the appropriate Condor library. It may not be suitable for use with all programs (e.g. when source code is not available), thus leading to restrictions in its use. The Vanilla universe, on the other hand, can be used with virtually all applications provided there are no license restrictions.

Jobs submitted to the Condor pool can be tracked using Condor job management utilities, such as *condor_q* and *condor_status*. These commands can be run with different options to get specific information, for example, the current status of a job, wall-clock time accumulated, the host where the job is running, network usage of job, etc. In addition, the status and availability of computing resources, as well as the nature of job allocations within these resources, can be monitored using the resource monitoring utility *CondorView*. In addition Condor, along with the Directed Acyclic Graph Manager (DAGMan) meta-scheduler, provides a facility to control computations where the input, output, or execution is dependent on one or more other computations. This is a very useful functionality provided within Condor for model runs that have dependencies. Additional information on Condor is available at the Condor Project home page.¹⁶

Computational resources associated with the Condor pool at Purdue University that are part of the Condor pool are used in this study. The Condor pool consists of Linux clusters (x86_64 and ia32 processors), Windows workstations (WinNT51) in computer labs, and a few Sun Solaris and MAC OSX nodes.³⁹ The Linux pool is the largest with more than 7000 processors (cores). These large Linux clusters are also part of the National Science Foundation (NSF) TeraGrid^{40,41} computational effort to provide open scientific discovery infrastructure for high-performance computational resources. Purdue University, a TeraGrid partner, is the largest provider of Condor cycles among academic institutions.

2.5. Procedure for making SWAT model simulations in the Condor framework

Watershed response predictions were needed from 2004–2028, representing 25 years of model runs for each BMP scenario. The stochasticity in future weather conditions was captured by generating 250 weather realizations for the period of simulation. Thus, the total number of model runs needed was 43,000 (172 management scenarios × 250 weather realizations for each scenario). Running the model on a standard Linux workstation took approximately 10 minutes/run. Thus, a simulation time of 7167 CPU hours (299 days) would be needed to complete all runs on a standard Linux system. Running the SWAT model in the Condor framework required model reconfiguration as explained in the subsequent sections.

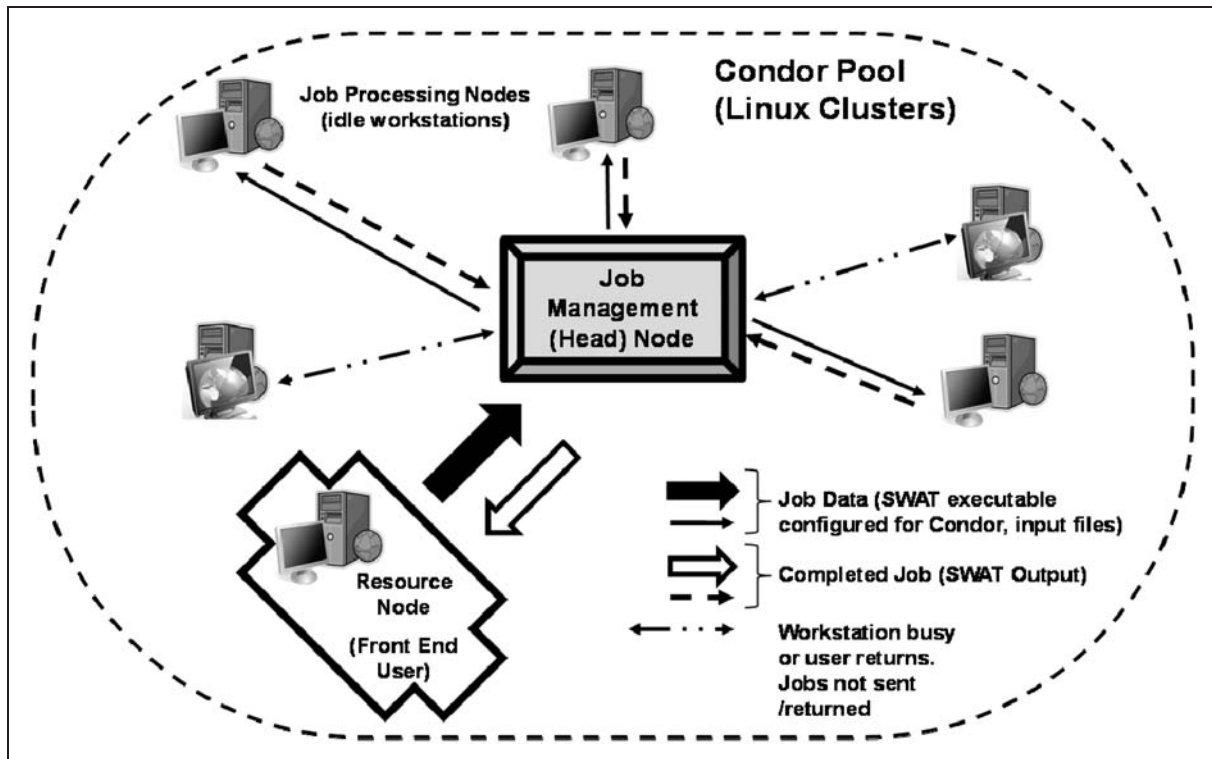


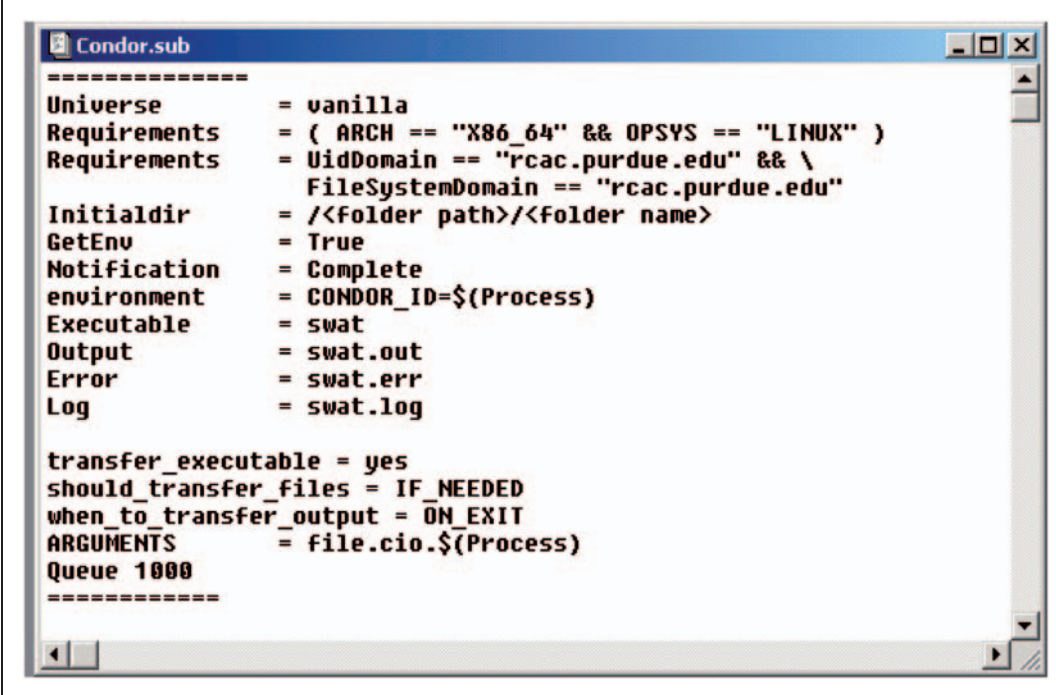
Figure 2. Depiction of Condor–Soil and Water Assessment Tool (SWAT) distributed computing system.

2.5.1. Modifications to the SWAT model code. The SWAT model has an open source code available at <http://www.brc.tamus.edu/swat/>. The model source code was modified to select the appropriate input files (the configuration file- file.cio and weather files including temperature and precipitation data) dynamically at run time. As mentioned earlier for each BMP scenario, 250 weather realizations were used; each input file with different data was suffixed with a number between 0 and 249 to account for the 250 scenarios. Accordingly, multiple HRU output file names were dynamically changed by the program.

By default, the SWAT outputs 68 different output variables of watershed response. Further, these HRU-level outputs contain average annual values of the same variables for each of the HRUs. Typically, a model user may be interested in only a few model outputs. The default output file size for each model run was greater than 100 MB. Storing all model outputs for further analyses was another challenge in this study. To reduce the size of the model output files and the subsequent data storage requirements, we modified the model in the options input file (file.cio) such that the model would output only those variables required for this study (flow, sediment yield, nitrogen and phosphorus (various forms), biomass, and yield). In addition, model codes were modified to remove information from the HRU output file that was not needed for our

analyses, including the land use and HRU area. While this information is pertinent to the output, it does not change from one run to another. With the exception of the HRU and sub-basin number, and the simulation month/year, reference to all such information was removed from the code. This information was re-created in a separate 'headers' file, which was then used with the HRU-level output during post-processing. The HRU and sub-basin numbers, and the month/year left within the HRU output file were used as checks to ensure that the 'header' and output data were matched accurately for subsequent data analysis. Adjustments were also made in the code to remove the average annual output as this information was not needed in further assessments. Further, all column titles were removed to facilitate easy reading of data during post-processing.

2.5.2. SWAT compilation in Linux. In order to run the SWAT in the Condor environment on Linux clusters, the first step was to port the Windows version of the SWAT program. This involved creation of a new instructions (make) file and fixing compilation errors. We used Intel Fortran compilers for the SWAT model code compilation. The program was thoroughly tested on the Linux cluster. The results were validated with results from an identical SWAT program in the Windows environment.



```

=====
Universe           = vanilla
Requirements       = ( ARCH == "X86_64" && OPSYS == "LINUX" )
Requirements       = UidDomain == "rcac.purdue.edu" && \
                    FileSystemDomain == "rcac.purdue.edu"
Initialdir         = /<folder path>/<folder name>
GetEnv             = True
Notification       = Complete
environment        = CONDOR_ID=$(Process)
Executable         = swat
Output             = swat.out
Error              = swat.err
Log                = swat.log

transfer_executable = yes
should_transfer_files = IF_NEEDED
when_to_transfer_output = ON_EXIT
ARGUMENTS          = file.cio.$(Process)
Queue 1000
=====

```

Figure 3. Condor submit file. Attributes give instructions to Condor on how to execute model runs.

Initial testing of the SWAT using the Condor framework in Vanilla Universe mode was done to plan for the large number of model runs. The runs were done in the Condor Linux pool using the Condor's job submission script. This script allows the user to select a particular platform (Linux 32 or 64 bits), processor type, and speed, or any specialized architecture, including the option to use a shared file system or to move the input files to the remote machine for execution. All the runs performed used NFS (Network File System) shared storage. This also reduced the network traffic, as otherwise the runs would have necessitated moving a large number of files back and forth, potentially leading to bottlenecks and relatively longer turnaround times.

2.5.3. Preparation for the Condor submission file. In order to perform the Condor runs, it was necessary to develop a Condor submission file (Figure 3), which is an input (text) file containing instructions to Condor on how to perform model runs. The Condor submit file comprises several attributes including the universe, which defines the environment in which runs are executed, the requirements, which define the desired machine architecture and operating system, as well as the file system domain, the environmental variables (environment), which comprises a set of variables that denote specific model runs, the executable, which gives the name of the executable file, and the arguments, which specify the input(s) that might otherwise have been supplied through the command line or by other

(non-automated) means. Additional information on this can be obtained from the Condor user manual.

A key attribute in the submit file is the queue, which specifies the number of different model runs to be executed. The term \$ (process) included with the environment and arguments attributes is a macro that supplies the values for the run numbers, in this case from 000 to 249. The submit file also includes the names for standard output, error, and log files, and other instructional and environment-specific attributes such as the 'transfer_executable' command, which tells Condor whether or not to transfer the executable to the remote (executing) machine.

2.5.4. Test of SWAT–Condor system runs. Once all configurations were set and inputs had been prepared, runs were conducted to test the SWAT–Condor system. Initial testing was conducted by submitting 10 runs comprising one scenario with 10 different weather realizations. Once these runs were complete, the execution time was noted. Average annual output at the watershed level from some of the runs were also checked against outputs from standard (non-Condor) runs to validate the results obtained from Condor batch runs. The system was further tested by submitting 250 runs, then 500 and 1000 runs for a single scenario. Observations were made regarding the timing and performance of the system for each of the sets of runs, as well as the availability and general performance of Condor computing resources.

3. Results

3.1. BMP performance analysis

The SWAT model simulations in the Condor framework enabled us to successfully make the large number of model runs needed for a comprehensive BMP performance analysis. Figure 4 shows an example result that was obtained from 43,000 SWAT model runs. The losses of total sediment, total N, mineral P, and total P under all 172 different BMP scenarios from the watershed considering the stochasticity in weather data are shown in this figure. The dashed horizontal lines represent the median and maximum amounts of water quality constituent losses due to weather stochasticity under the current watershed management conditions (baseline). The black vertical boxes represent the first and third quartile losses of water quality constituents and the light gray lines represent the range of the losses. It should be noticed that the performance of the edge-of-field filter strip in the SWAT model is based on simple empirical functions of filter width (FILTERW). The filter strip trapping efficiency for sediment and nutrients due to the filter strip are calculated as⁴² $\text{trap} = 0.367(\text{FILTERW})^{0.2967}$, resulting in a large trapping efficiency for a 30 m filter width. It is evident from Figure 4 that weather stochasticity presents a large variability in losses of sediment, N, and P from the watershed. For the majority of the BMP scenarios, the median losses of water quality constituents were less than the losses under the current watershed management conditions, indicating that these BMPs can be expected to decrease NPS losses of sediment, N, and P from pasture areas. However, under extreme weather conditions, the maximum constituent losses for some BMP scenarios, especially for overgrazing management, are even greater than the maximum losses under the baseline scenario, indicating that overgrazing management should be avoided to see any improvement in the water quality. Uncertainty in baseline losses of sediment, total N, mineral P, and total P from pasture HRUs, as indicated by a coefficient of variation (CV) was 0.95, 0.4, 0.58, and 0.67, respectively. Similarly, ranges of CV for predicted losses of sediment, total N, mineral P, and total P from pasture HRUs due to weather stochasticity were 0.79–1.28, 0.56–1.11, 0.61–0.84, and 0.7–0.86, respectively. A large CV for constituent losses under baseline conditions indicate that weather uncertainty must be considered when evaluating BMP performance or when setting up goals for water quality improvement due to BMP implementation. More details of the reduction rates of nutrient losses from those 171 BMPs and the amounts of nutrient losses possibly contributed by weather uncertainty can be found in Chaubey et al.¹⁸

3.2. BMP effectiveness assessment

SWAT model runs using Condor enabled us to quantify interactions among BMPs and to evaluate whether they had synergistic or contrasting effects when implemented concurrently in the watershed. For example, Figure 5–7 show the interactions among grazing (no grazing, optimum grazing, and over grazing), non-amended poultry litter application rates in the watershed (1, 1.5, 2, 2.5, and 3 tons/ac), and timing of poultry litter application (spring, summer, and fall) on total sediment, N, and P losses from the pasture areas and from the entire watershed, respectively. A total of 27 BMP scenarios were selected and grouped by different grazing management for representing the interactions among different BMP categories. Each dot represents the median of simulation results for each selected BMP scenario under various weather conditions, and the dashed lines show constituent losses when no poultry litter is applied. An increase in intensity of grazing management (from no grazing to overgrazing management) resulted in increasing constituent losses in the watershed. Both N and P losses from pasture areas increased with an increase in litter application rates. While N losses were greatest for fall litter application for all grazing management, P losses were not sensitive to litter application timing for no grazing and optimum grazing management. The interaction effects between litter application timing and grazing management on P losses indicated that low-intensity grazing management had greater impacts on P losses than litter application timing. When all 171 BMP scenarios were compared together, buffer strips were the most important BMPs affecting the losses of total N from the pasture areas, and litter application timing, buffer strips, and grazing management were the three most important BMPs affecting the losses of total P.¹⁸

3.3. Challenges encountered in the Condor applications of the SWAT model

While Condor enabled us to efficiently perform thousands of SWAT simulations to comprehensively evaluate BMP performance, their interactions and the impact of weather stochasticity on water quality, several challenges were encountered during this application. We have briefly outlined these challenges to benefit the modeling community, who may be interested in similar applications.

3.3.1. Submitting a large number of jobs at once may lead to system overload and job termination. At the beginning, 1000 weather data simulations were planned for each BMP scenario.

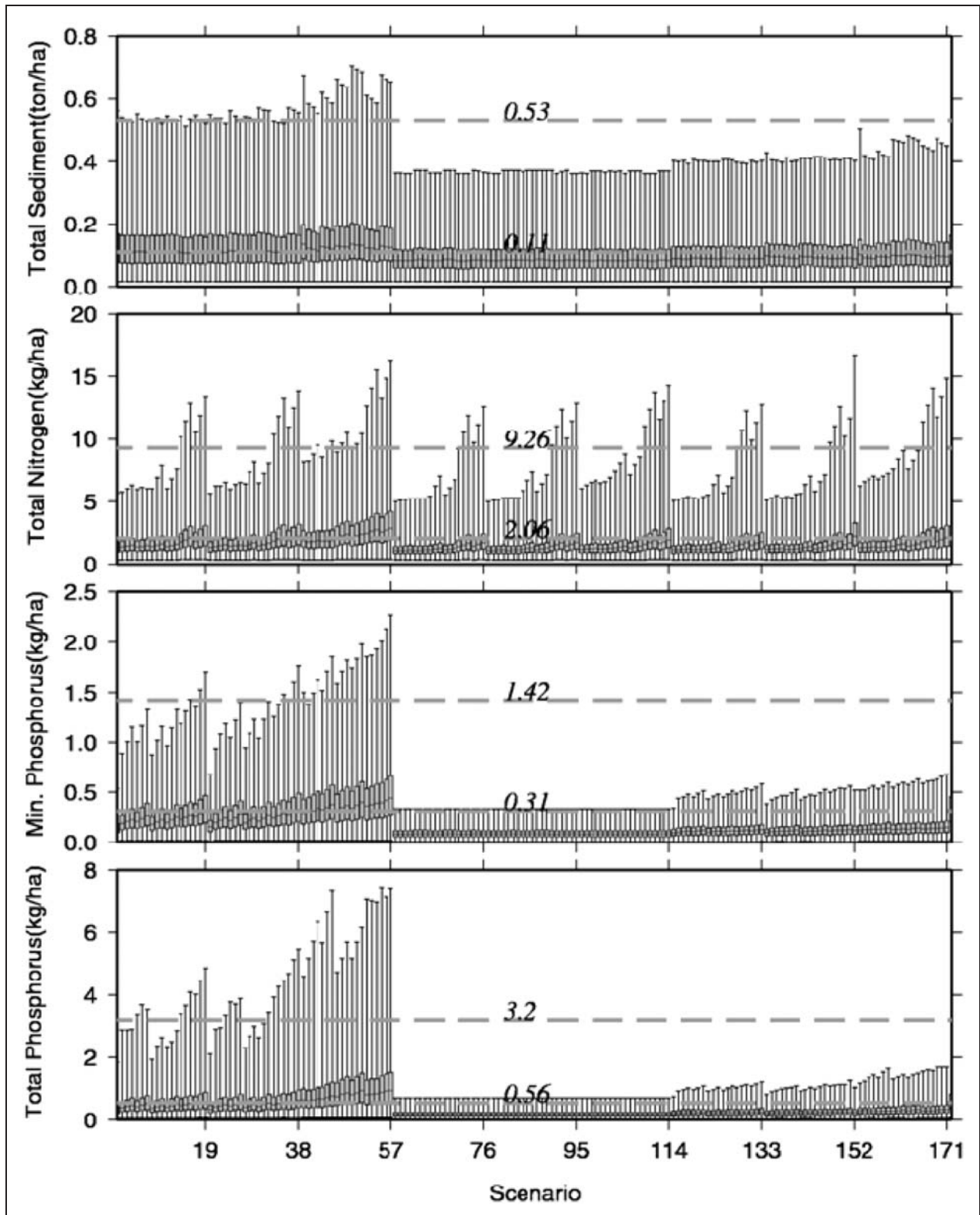


Figure 4. Performance of 172 best management practice scenarios on total sediment, total nitrogen, mineral phosphorus, and total phosphorus losses in the watershed under various weather conditions. Horizontal lines represent median and maximum constituent losses under the current watershed conditions (baseline scenario) due to stochasticity in weather conditions.

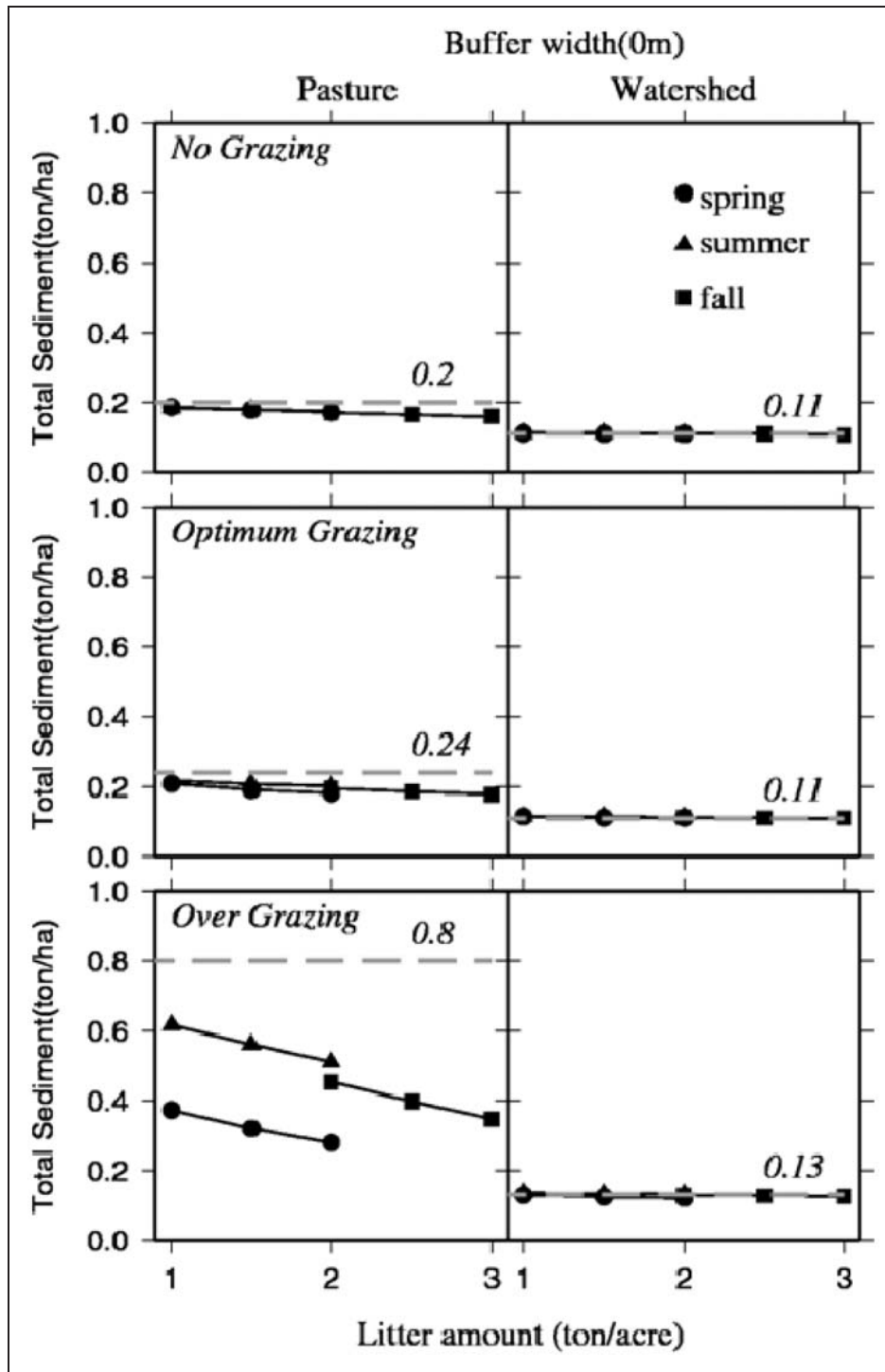


Figure 5. Performance of selected 27 best management practice scenarios with no buffer zone on total sediment loss from pasture areas and from the entire watershed.

However, with 1000 simulation jobs submitted to Condor, it took 10–15 days to complete these jobs because of the policy to lower a user’s priority when jobs requesting more than 200–300 processors are submitted at once or if other jobs are waiting in the Condor queue. In this application, submitting 250 jobs for three

scenarios (i.e. a total of 750 jobs) resulted in a considerably shorter completion time than submitting 1000 jobs for only one scenario. Table 2 shows the summary of Condor performance for various job submissions. Usually, we noticed faster turnaround times during weekends, indicating that the jobs were getting a

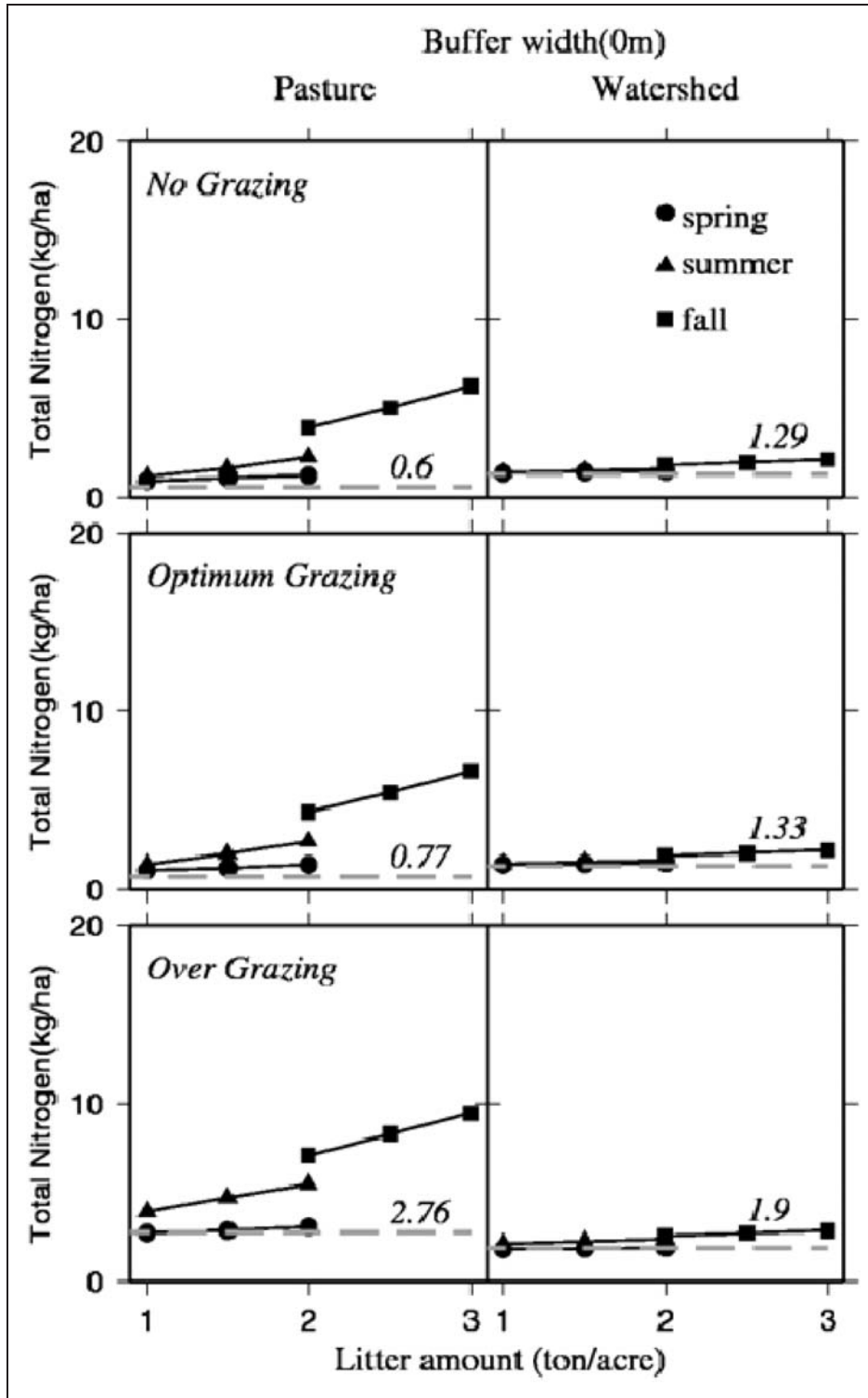


Figure 6. Performance of selected 27 best management practice scenarios with no buffer zone on total nitrogen loss from pasture areas and from the entire watershed.

higher priority when fewer users were submitting Condor jobs. In this case, there was no quick fix and the jobs were completed based on the priority assigned to them during the submission.

3.3.2. Making a large number of SWAT runs in the Condor environment requires a large data storage space. In many applications, all the information printed in the SWAT output files may not be needed,

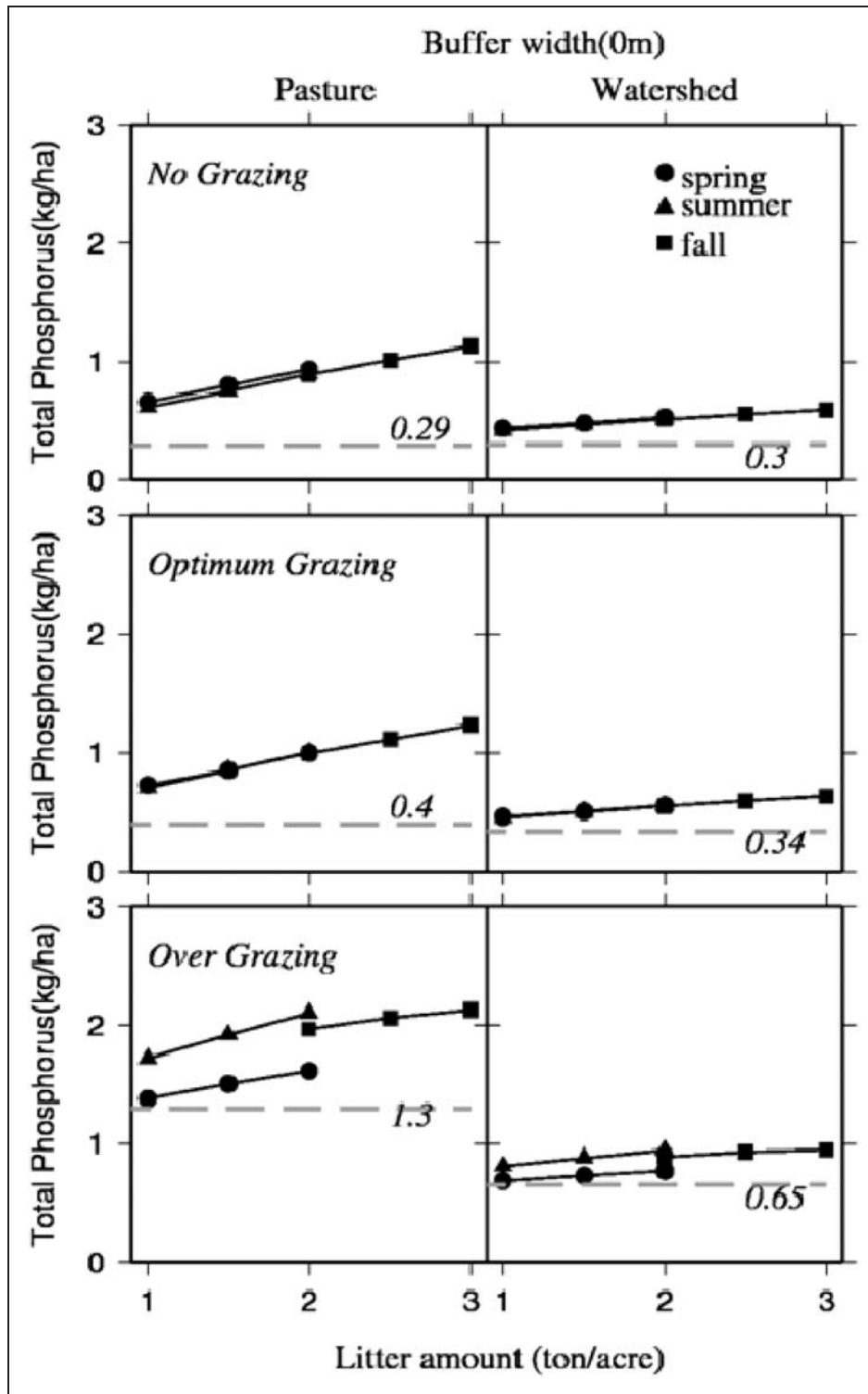


Figure 7. Performance of selected 27 best management practice scenarios with no buffer zone on total phosphorus loss from pasture areas and from the entire watershed.

and thus a post-processor can be written or SWAT code can be modified to only print the data that are needed, thus effectively reducing the total data storage requirements. For this study, we first modified the

SWAT code to remove text headers and redundant (with respect to this study) columns, thus printing only the required output. We then developed a post-processor in MATLAB[®] that reduced each HRU

Table 2. Summary of Condor performance for different job submissions

# of jobs submitted	Time to completion	Comments
1 scenario, 1000 jobs total	10–15 days	200–300 jobs done, 700–800 jobs idled Forceful release required to run and complete idled jobs Faster on evenings, weekends (fewer users)
3 scenarios, 250 jobs each	2–15 hrs	More efficient – the three scenarios distributed to three different machines Speed still influenced by number of users
172 scenarios, 250 jobs each, 43,000 jobs total	6–18 days	A maximum of 2500 jobs for 10 scenarios can be completed in one research account per day due to the limits of file numbers and size. Submission from more research accounts would facilitate the completion of 172 scenarios.

output file from 57 MB to 7 MB and each watershed output file from 2.3 MB to 15 kB, respectively. In addition, we could not store all the original output files in the project directory that handled all Condor runs, because they required more than 2.5 TB of storage space. We thus moved these files to an alternate archival directory.

3.3.3. Post-processing of output files was time consuming. The BMP scenarios were first generated on a local Windows machine and each BMP scenario folder contained more than 7000 SWAT input files. In order to facilitate the transfer of SWAT input files to the remote TeraGrid resources, each BMP scenario folder was compressed before file transfer and then uncompressed. It took approximately 10 minutes to uncompress this folder. Three BMP scenarios, a total of 750 SWAT runs, were submitted as a batch job to the TeraGrid–Condor system. These 750 SWAT runs normally completed within 2 hours. After the SWAT runs of each BMP scenario were completed, the three primary output files (output.hru, output.std, and out.out) were transferred back to the host machine (PC). Due to the different sizes of output files, downloading time for 250 output.hru (size: 57MB/output.hru), 250 output.std (size:0.4MB/output.std), and 750 out.out (size: 2.2MB/out.out) files was 62, 4, and 12 minutes, respectively. Post-processing of output.hru files using MATLAB code was the most time-consuming step, which was used to obtain the monthly and annual water quality values from different land uses at the sub-basin level. Due to the complexity of the post-processing step it was done on local computers, instead of faster TeraGrid computers. Therefore, it took 13 hours to complete the post-processing for one BMP scenario and approximately two weeks to complete all BMP scenarios using five computers simultaneously. Further modification of the SWAT code is recommended to provide flexibility

in generating only the required output, thus reducing considerably the file size and possibly eliminating the need for post-processing.

4. Summary and conclusion

Assessment of impacts of watershed management decisions on water quality frequently requires running complex watershed models. However, the computational time required to evaluate uncertainty from various input data sets or a decision matrix can be very high and may not be feasible if only a few computers are available to run hundreds of thousands of model simulations. In this study we integrated the SWAT model into the Condor environment running on TeraGrid, a large network of computers. More than 43,000 model runs were performed to evaluate impacts of various BMPs and weather uncertainty on water quality. These model runs were efficiently performed in a few days using the high-throughput computing Condor framework. If we had to complete these runs on single desktop computer workstations (Intel Xeon 2.8-3.5 GHz dual core processors, 2.5 GB of RAM, 250+ GB disk), the same runs could have taken 2.5–3.3 years to complete. With greater emphasis to evaluate the efficacy of various BMPs and to fully quantify parameter and output uncertainty, there is a need to perform such evaluations efficiently. The approach demonstrated in this study can be used in other watersheds or other complex models and can potentially reduce the model run time considerably.

Acknowledgements

We acknowledge the support by the NSF-TeraGrid program to provide access to the Condor network for model simulations and the technical support by the Purdue University High Performance Computing Center. Comments provided by the three anonymous reviewers greatly improved an earlier version of this manuscript.

Funding

This study was supported by the USDA-CSREES under the CEAP [project number 2005-48619-03334].

Conflict of interest statement

None declared.

References

- Migliaccio KW and Srivastava P. Hydrologic components of watershed-scale models. *Trans ASABE* 2007; 50: 1695–1703.
- Arabi M, Govindaraju RS and Hantush MM. A probabilistic approach for analysis of uncertainty in the evaluation of watershed management practices. *J Hydrol* 2007; 333: 459–471.
- Green CH, Arnold JG, Williams JR, Haney R and Harmel RD. Soil and water assessment tool hydrologic and water quality evaluation of poultry litter application to small-scale subwatersheds in Texas. *Trans ASABE* 2007; 50: 1199–1209.
- Gitau MW, Veith TL, Gburek WJ and Jarrett AR. Watershed level best management practice selection and placement in the town brook watershed, New York. *J Am Water Resour Assoc* 2006; 42: 1565–1581.
- Tong STY and Naramngam S. Modeling the impacts of farming practices on water quality in the little Miami River Basin. *Environ Manage* 2007; 39: 853–866.
- Ferreira VA, Weesies GA, Yoder DC, Foster GR and Renard KG. The site and condition specific nature of sensitivity analysis. *J Soil Water Conserv* 1995; 50: 493–497.
- Arnold JG, Srinivasan R, Muttiah RS and Williams JR. Large area hydrologic modeling and assessment -Part 1: Model development. *J Am Water Resour Assoc* 1998; 34: 73–89.
- Gitau MW, Srivastava R and Chaubey I. Watershed response modeling in Arkansas priority watersheds: Experience with SWAT autocalibration An ASABE Meeting Paper: 2007 ASABE Annual International Meeting, Minneapolis, MN. ASABE. St. Joseph, MI. Paper Number: 072171:2007.
- Zhang X, Srinivasan R and Van Liew M. Multi-site calibration of the SWAT model for hydrologic modeling. *Trans ASABE* 2008; 51: 2039–2049.
- Bekele EG and Nicklow JW. Multi-objective automatic calibration of SWAT using NSGA-II. *J Hydrol* 2007; 341: 165–176.
- Shirmohammadi A, Chaubey I, Harmel RD, et al. Uncertainty in TMDL models. *Trans ASABE* 2006; 49: 1033–1049.
- Shirmohammadi A, Chu TW and Montas HJ. Modeling at catchment scale and associated uncertainties. *Boreal Environ Res* 2008; 13: 185–193.
- Migliaccio KW and Chaubey I. Spatial distributions and stochastic parameter influences on SWAT flow and sediment predictions. *J Hydrol Eng* 2008; 13: 258–269.
- Maringanti C, Chaubey I, Arabi M and Engel B. A multi-objective optimization tool for the selection and placement of BMPs for pesticide control. *Earth Syst Sci Data Discuss* 2008; 5: 1821–1862.
- Litzkow MJ, Livny M and Mutka MW. Condor – A hunter of idle workstations. *IEEE* 1988: CH2541-1/88/0000/0104\$01.00.
- Condor Team. ‘Condor Project’, University of Wisconsin-Madison, <http://www.cs.wisc.edu/condor> (1990, accessed 20 July 2007).
- Epema DHJ, Livny M, van Dantzig R, Evers X and Pruyne J. A worldwide flock of Condors: Load sharing among workstation clusters. *Future Gener Comp Syst* 1996; 12: 53–65.
- Chaubey I, Chiang L, Gitau MW and Mohamed S. Effectiveness of BMPs in improving water quality in a pasture dominated watershed. *J Soil Water Conserv* 2010; 65: 424–437.
- Chiang L, Chaubey I, Gitau MW and Arnold JG. Differentiating impacts of land use changes from pasture management in a CEAP watershed using SWAT model. *Trans ASABE* 2010; 53: 1569–1584.
- Edwards DR and Daniel TC. Effects of poultry litter application rate and rainfall intensity on quality of runoff from fescuegrass plots. *J Environ Qual* 1993; 22: 361–365.
- Edwards DR, Daniel TC, Scott HD, Moore PA, Murdoch JF and Vendrell PF. Effect of BMP implementation on storm flow quality of two northwestern Arkansas streams. *Trans ASAE* 1997; 40: 1311–1319.
- Edwards DR, Daniel TC, Scott HD, Murdoch JF, Habiger MJ and Burks HM. Stream quality impacts of best management practices in a northwestern Arkansas basin. *Water Resour Bull* 1996; 32: 499–509.
- Sharpley AN, Chapra SC, Wedepohl R, Sims JT, Daniel TC and Reddy KR. Managing agricultural phosphorus for protection of surface waters -Issues and options. *J Environ Qual* 1994; 23: 437–451.
- Arnold JG, Allen PM and Bernhardt G. A comprehensive surface-groundwater flow model. *J Hydrol* 1993; 142: 47–69.
- White KL and Chaubey I. Sensitivity analysis, calibration, and validations for a multisite and multivariable SWAT model. *J Am Water Resour Assoc* 2005; 41: 1077–1089.
- Gitau MW, Gburek WJ and Bishop PL. Use of the SWAT model to quantify water quality effects of agricultural BMPs at the farm-scale level. *Trans ASABE* 2008; 51: 1925–1936.
- Gassman PW, Reyes MR, Green CH and Arnold JG. The soil and water assessment tool: Historical development, applications, and future research directions. *Trans ASABE* 2007; 50: 1211–1250.
- Olivera F, Valenzuela M, Srinivasan R, et al. ArcGIS-SWAT: A geodata model and GIS interface for SWAT. *J Am Water Resour Assoc* 2006; 42: 295–309.
- Miller SN, Semmens DJ, Goodrich DC, et al. The Automated Geospatial Watershed Assessment tool. *Environ Modell Softw* 2007; 22: 365–377.
- Green CH and van Griensven A. Autocalibration in hydrologic modeling: Using SWAT2005 in small-scale watersheds. *Environ Modell Softw* 2008; 23: 422–434.

31. Gollamudi A, Madramootoo CA and Enright P. Water quality modeling of two agricultural fields in southern Quebec using SWAT. *Trans ASABE* 2007; 50: 1973–1980.
32. Jha MK, Gassman PW and Arnold JG. Water quality modeling for the Raccoon River watershed using SWAT. *Trans ASABE* 2007; 50: 479–493.
33. Quansah JE, Engel BA and Chaubey I. Effects of poultry litter application rate and rainfall intensity on quality of runoff from fescuegrass plots. *Trans ASABE* 2008; 51: 1311–1321.
34. NRCS. 'Soil Data Mart,' Natural Resources Conservation Service, United States Department of Agricultural, <http://soildatamart.nrcs.usda.gov/> (2001, accessed 20 December 2001).
35. Sharpley AN and Williams JR. EPIC-erosion productivity impact calculator, 1.model documentation. U.S. Department of Agriculture, Agricultural Research Service. *Tech. Bull* 1990; 1768.
36. UAEX (University of Arkansas Cooperative Extension Service). Forage and Pasture Forage Management Guides. 'Self-Study Guide 5: Utilization of Forages by Beef Cattle', http://www.aragriculture.org/forage_pasture/Management_Guides/Forages_Self_Help_Guide5.htm (2006, accessed August 2007).
37. Gitau MW. A quantitative assessment of BMP effectiveness for phosphorus pollution control: The Town Brook Watershed, NY. *PhD Dissertation*. University Park, PA: The Pennsylvania State University, 2003.
38. Boer R. Resource management in the Condor system. *Master's Thesis*. Delft University of Technology, 1996.
39. RCAC (Rosen Center for Advanced Computing). 'BoilerGrid', <http://www.rcac.purdue.edu/userinfo/resources/boilergrid/> (1987). Accessed August 2007.
40. TeraGrid. <http://www.teragrid.org/> (2001 accessed August 2007).
41. Walker E, Gardner JP, Litvin V and Turner EL. Personal adaptive clusters as containers for scientific jobs. *Cluster Comput* 2007; 10: 339–350.
42. Neitsch SL, Arnold JG, Kiniry JR, Srinivasan R and Williams JR. *Soil and Water Assessment Tool input/output file documentation, version 2005*. Temple, TX: Blackland Research Center, USDA Agricultural Research Service, 2004.

Margaret W Gitau is an assistant professor in Biological and Agricultural Systems Engineering at Florida A&M University, Tallahassee, FL, USA. She also serves as an Associate Editor for the Transactions of the American Society of Agricultural and Biological Engineers and as a reviewer for various internationally recognized journals.

Li-Chi Chiang is a student researcher at the US Environmental Protection Agency, National Exposure Research Laboratory – Environmental Science Division, Las Vegas, NV, USA. She earned her Doctorate at Purdue University, Department of Agricultural and Biological Engineering, West Lafayette, IN, USA.

Mohamed Sayeed obtained his PhD in 2004 from North Carolina State University in Civil Engineering with a focus in computational science and high-performance computing. After his PhD he worked as a research scientist at Purdue University. Currently he is working as a director of MWM ventures LLC, a technology venture capital firm of which he is also one of the co-founders.

Indrajeet Chaubey is an Associate Professor in the Departments of Agricultural and Biological Engineering, and Earth and Atmospheric Sciences at Purdue University. He was the project director of the Arkansas CEAP. He has several active research projects funded by the US Environmental Protection Agency, US Department of Agriculture (USDA), NSF, and US Department of Energy (DOE). Currently he is leading several projects funded by the USDA and DOE to evaluate the hydrologic/water quality impacts and sustainability of biomass production for advanced biofuels in Midwest and Southeast USA.