High performance computing application to address non-point source pollution at a watershed level

Chetan Maringanti,
Graduate Student, Department of Agricultural and Biological Engineering, Purdue University, West Lafayette, IN. 47906, email: cmaringa@purdue.edu

Indrajeet Chaubey,
Associate Professor, Department of Agricultural and Biological Engineering and Department of Earth and Atmospheric Sciences, Purdue University, West Lafayette, IN. 47906, email: ichaubey@purdue.edu

Abstract. Best management practices (BMPs) have been proven to effectively reduce the Nonpoint source (NPS) pollution loads from agricultural areas. BMP selection and placement problem needs to be addressed for the placement of BMPs in a watershed at a field scale that would achieve maximum pollution reduction subjected to minimum cost increase for the placement of BMPs in the watershed through optimization techniques. The BMP selection problem is linked with a hydrologic and water quality simulation model to estimate the pollution loads at various locations in the watershed. However, the optimization techniques get very complicated as the size of the watershed gets larger as the design variables increase proportionately. Also, the BMP optimization is linked to the watershed level hydrologic and water quality simulation mode and this makes the problem require very high amount of computation time for the simulations to get to a near optimal set of BMP placement in the
watershed. In this regard, the high performance computing (HPC) helps a great deal by providing a platform for running the simulations on a multitude of processors rather than a single processor. Unix serves as the de-facto operating system for most of the HPC applications as the architecture allows to run simulations that take a long time to run in much lesser time when compared to the Windows OS machines. In the present work we have used the Soil and Water Assessment Tool (SWAT) model in combination with a genetic algorithm (GA) optimization technique NSGA-II to optimally select and place BMPs at a watershed scale. We have tested the application of the optimization algorithm to reduce sediment, nitrogen, and phosphorus from the L’Anguille River Watershed, Arkansas and pesticide from Wildcat Creek Watershed, Indiana. The SWAT model was compiled in Unix and the optimization routine for BMP selection and placement was coded as an objective function in NSGA-II program in C language. The optimization model was compiled in Unix to optimize the various BMPs that can be placed at a field scale, which is represented as a Hydrologic Response Unit (HRU) in SWAT.

**Keywords.** High performance computing, BMP, optimization, nonpoint source pollution
Introduction

Best management practices (BMP) selection and placement problem consists of finding an optimum solution for BMP placement in a watershed that will achieve the two most important objectives of maximum pollution reduction with a minimum net cost increase for the implementation of BMPs. It is required to optimize the two objectives with the variables being the BMPs at every farm level in the watershed. Optimization algorithms such as genetic algorithms are widely used for BMP selection and placement in a watershed. Most of the optimization models developed are linked dynamically with a watershed level simulation model to estimate the pollutant loads at a watershed scale during the optimization [Arabi et al., 2006; Bekele and Nicklow, 2005; Srivastava et al., 2002]. Presence of the dynamic linkage in the BMP selection and placement models restrict their application to relatively smaller watersheds. Also, the number of possible solutions to be searched during optimization increases exponentially with the number of BMPs considered for implementation due to which a vast number of BMPs were not considered. This problem was addressed through development of a novel methodology that utilizes a database, called as BMP tool, to estimate the pollution effectiveness of any given BMP and therefore removes the requirement of a dynamic linkage with the watershed simulation model during the optimization [Gitau et al., 2006; Maringanti, 2008]. Also, the model for optimization can be run for multiple pollutants and for multiple years of consideration independent to each other and this is a slightly computationally intensive job. More information regarding the model and the method used can be obtained from Maringanti et al., [2009]. In order to validate the applicability of the solutions obtained after optimization using the BMP tool, some solutions need to be picked up from the Pareto-optimal front and evaluated through
the watershed model to check if similar pollution reductions were obtained. The evaluation of these solutions using the watershed model is again a computationally time intensive job.

High performance computing (HPC), that consists of super computers and cluster of processors are used to solve problems that are highly computation intensive. Multiple jobs that are required to accomplish a particular task are distributed to make them run in parallel among multiple processors and therefore decreasing the overall wall clock time for execution. HPC has been used extensively in calibrating hydrologic [Confesor and Whittaker, 2007] and environmental modeling [Vrugt et al., 2006]. As the number of processors increase, the overall computation time is reduced in a range of 50% - 80% [Vrugt et al., 2006].

HPC would greatly help reduce the computation time required to run replicates of the optimization algorithm for multiple pollutants and multiple average years of concern. Also, the computation time required to evaluate the watershed model the solutions picked from the BMP tool output would be highly reduced using HPC.

Objectives

The overall goal of this study is to evaluate the performance of HPC based on a Unix based cluster when compared to the performance of a single processor in windows. In order to achieve this overall goal, the following tasks are performed

1) The BMP optimization tool based on a genetic algorithm is run for nitrogen, phosphorus, and sediment along with 5 different averages of pollutant loads (5yr, 10yr, 15yr, 20yr, and 25 yr) in Lincoln Lake Watershed, Arkansas.
2) The validation of the solutions obtained from the optimization of the BMP optimization tool are used to evaluate the pollutant loads using a watershed model to estimate nitrogen, phosphorus, and sediment in L’Anguille River Watershed; and pesticide in Wildcat Creek Watershed.

Theoretical Background

Multi-objective genetic algorithm (MOGA)

Genetic algorithms are heuristic based search algorithms, based on Darwin’s theory of evolution, that function well in searching optimum solutions when the search space gets large in the variable and objective function domain. As the BMP selection and placement problem requires two objectives (reduction in pollutant loads and reduction in net cost for placement of BMPs) to be minimized simultaneously, it is required to use a multi-objective genetic algorithm (MOGA) for optimization. Non-dominated sorted genetic algorithm (NSGA-II) (Deb, 1999, 2001; Deb et al., 2002) is one such MOGA that can search a large number of variables and objective functions space to find an optimal solution. The overall computation complexity of the algorithm is \(O(MN^2)\), which usually is \(O(MN^3)\) for most of the evolutionary techniques (Deb et al., 2002): where \(O\) is ‘order of’, \(M\) is the number of objective functions, and \(N\) is the population size. Non-domination sorting, crowding distance, and elite sets are important properties of the NSGA-II algorithm that ensure that the solutions that are obtained during the optimization have a good spread in the objective function space and the optimization is ensured to reach a Pareto-optimum (Figure 1). This also ensures that the algorithm does not undergo a premature convergence. Crossover and mutation are other genetic operations that are important to generate the future
offspring based on the parental genetic information and create genetic modification to reach convergence respectively.

Watershed model (Soil and Water Assessment Tool)

The watershed model used in this study was Soil and Water Assessment Tool (SWAT) model. The SWAT is a process based distributed parameter watershed scale simulation model designed for use in gauged as well as ungauged basins to simulate long term effects of various watershed management decisions on hydrology and water quality response (Arnold et al., 1998). The SWAT model divides the watershed into subwatersheds or subbasins based on the outlets selected within the watershed by the user and further divides subbasins into land areas, called hydrologic response units (HRUs), based on land use, management, and soil properties. SWAT performs most of the calculations at HRU level, which are then delivered into the stream present in the particular subbasin and then routed to the downstream subbasin.

The input data needed by the model are related to watershed physical characteristics, climate, plant growth, reservoir data (if any), and management practices at HRU level. The typical GIS data needed by the model are digital elevation map (DEM), soil, and land use maps. The climatic input data required by SWAT are precipitation, temperature, solar radiation, relative humidity, and wind on a daily scale from multiple climatic gauge locations. Agricultural activities can be given as input to the model by modifying the management files. SWAT simulates the flow, nutrients, sediment, and chemicals at the subbasin or HRU level. Surface runoff is computed using a modification of the SCS curve number technique (Soil Conservation Service, 1972). Modified Universal Soil Loss Equation (MUSLE) (Williams, 1975) is used by the SWAT model to estimate the soil erosion and sediment yield in the watershed. Nitrogen and phosphorus are
applied through fertilizer, manure or residue application. Bulk of phosphorus transport takes
place in the adsorbed form to the sediment whereas nitrogen is highly dissolvable in water and
therefore transported mainly through surface/subsurface flow. The important feature in SWAT is
that it aids in modeling the various BMPs (structural and management based) by changing
appropriate parameters in the input files of the model. This feature is utilized in the development
of BMP tool which estimates the effectiveness of BMPs for a particular pollutant reduction.
More details about the development of BMP tool can be obtained from [Maringanti, 2008].

**High performance computing facility**

The Steele supercomputing facility at Purdue University consists of a cluster that is made up of
893 Dell dual quad-core computer nodes with an estimate peak performance of 60 teraflops. The
machines run on Red Hat Linux operating system with various computational packages installed
on each machine. Steele supercomputing facility is the largest among the universities located in
the Midwest region of the United States and ranks 104 in the TOP500 list of global
supercomputers. The parallel simulations of the various model runs were run on the Steele
cluster from which our research group owns 4 nodes (32 processors).

**Study area and data available**

Lincoln Lake watershed (LLW) is located within the Illinois River Basin located in Northwest
Arkansas. The watershed has a total area of 32 km² with a mixed land use consisting of pasture,
forest, urban residential, urban commercial and water representing 35.84%, 48.6%, 11.9%,
1.54% and 2.2% of the watershed area, respectively (Figure 1). The watershed contains more
several hundreds of chicken houses. The 303 (d) lists Illinois River an impaired water body with
phosphorus being the main cause of impairment.
L’Anguille river watershed (LRW) is located in the Mississippi delta region of east Arkansas (Figure Error! No text of specified style in document.). The watershed covers an area of 2520 km$^2$ and drains the entire stretch (157 km) of the L’Anguille River. LRW is predominantly an agricultural watershed; with more than 60% of the watershed in arable land. The main crops grown in the watershed are rice (26%) and soybean (46%) and represent most (~95%) of the agricultural land in the watershed (Error! Reference source not found.). The L’Anguille River was included in the list of impaired water bodies by the Arkansas Department of Environmental Quality (ADEQ) in 1998 (ADEQ, 2002). Excessive sediment and nutrients are the main source of impairment of the river (ADEQ, 2002).

Wildcat Creek watershed (WCW) (Figure Error! No text of specified style in document.) (USGS 8 digit HUC 05120107) with a drainage area of 1956 km$^2$, located in north central Indiana was used for testing the optimal BMP selection and placement. The watershed is predominately agricultural with 74% row crops (36% Soybean, 38% Corn), 21% pasture, and 3% urban area (NASS, 2001). A high pesticide (atrazine) level from the agricultural areas has degraded the water quality of many watersheds in Indiana (Homes et al., 2001). Various NPS pollution reduction projects are being undertaken in the watershed through funds from the clean water act section 319. These funds, to effectively be utilized to reduce NPS pollutants, need to be combined with a BMP optimal selection and placement tool to choose the best economical and ecological alternative. The proposed multi-objective optimization methodology can be applied for the selection and placement of BMPs in Wildcat Creek Watershed for pesticide (atrazine) reduction from areas planted with corn.
**Methodology**

Figure 5 describes the methodology that is adopted during the computation process. Nitrogen, phosphorus, and sediment are the three pollutants considered for the LRW. The SWAT model that is used as a pollutant load estimator was simulated for 25 years. Five different annual averages (for 5, 10, 15, 20, and 25 years) were considered to study how the annual pollutant loads are changed due to the averaging process. These 5 averages and 3 different pollutants of concern require 15 different BMP selection and placement model runs to obtain the respective Pareto-optimal solutions after the optimization. These 15 model runs were made to run on Windows machines using a Windows compiled executable and on Steele supercomputer using a Unix compiled executable. The multiple jobs that are required for evaluating the optimization model are distributed into multiple processors on Steele to generate the final Pareto-optimal fronts that provide a range of solutions for the two objective functions.

Solutions from the final Pareto-optimal front are picked ensuring that the entire range is covered. N, P, and sediment optimization solutions were used from the LRW and pesticide solutions were used from the WCW. The variables that represent the BMPs to be implemented in each of the HRU are used to modify the input SWAT files and simulate the SWAT model to get the pollutant loads considering the in stream routing processes. The SWAT model used in the simulations was compiled in Unix and the runs are made on the Steele for different possible SWAT input configurations corresponding to the implemented BMPs.
Results and discussion

Table 1 illustrates how the SWAT model computation times vary for the various watersheds considered in the study using both Windows and Unix compiled versions. It is observed that the Unix version of SWAT outperforms the Windows counterpart by being more than 20 times computationally efficient.

Output from the SWAT model is used to develop the BMP tool and as an input into the NSGA-II and the optimization run is made with a population size of 800 and run for 20,000 generations. The optimization run took more than 12 hours to run on a Windows machine and less than 1 hour to run on a Unix machine. The 15 optimization runs required were distributed into 15 different processors and therefore the entire computation required less than an hour to get the final optimal solutions. If the same analysis was performed on a single Windows machine, it would have taken more than 7 days to complete the optimization. Figure 6 describes the progress of the Pareto-optimal front with the increasing generation number for sediment reduction. It is observed that the optimization results provide a range of solutions for the pollution reduction from the baseline. Similar analysis was performed for phosphorus and nitrogen.

Figure 7 presents the progress of the optimization front as the generation number increases for pesticide control in LRW. It is observed that at the end of 5000 generations, the solutions are spread well to provide a range of solutions for the selection and placement of BMPs in the watershed for pesticide control. 8 different solutions spanning the entire spread of the Pareto-optimal front at the end of generations 10, 20, 50, 5000 are selected and evaluated using the SWAT model to estimate the pollutant loads. These 32 different model runs were independent
of each other and therefore were evaluated in 32 different processors. This reduced the computation time from 128 minutes on a single Windows machine to less than a minute when run parallel on Steele.

Figure 9 presents the progress of the Pareto-optimal front as the generation number increases and the Pareto-optimal front at the end of the optimization for P, N, and sediment reduction in LRW. It is again observed that the Pareto-optimal front at the end of the final generation provides a range of alternative solutions for selection and placement of BMPs that can be chosen based on the required criteria for pollution reduction or available funds for BMP implementation. Figure 10 represents the SWAT simulated pollutant loads with the corresponding costs for the solutions 20 solutions picked from the range of the Pareto-optimal fronts (Figure 9) for generations 10, 40, 100, 1000, and 5000. The 20 picked solutions from the selected generations were set up to run on 20 different processors. The overall computation time to run these 100 different solutions was less than 5 minutes which would have been at least 16 hours when evaluated on a single CPU running Windows.

Conclusion

It is observed from the study that high performance computing (HPC) helps reducing the overall computation time to simulate multiple tasks by distributing them to run in parallel on multiple processors. The overall computation time to run the optimization model for BMP selection and placement was reduced by more than 90% by performing the runs in parallel on a supercomputer in contrast to a single processor run on Windows. Also, the overall computation time required to run the solutions picked from the simplified BMP tool based optimization technique to run the
SWAT model was also reduced by more than 90% because of the high performance supercomputing alternative chosen when compared to single machine Windows run. The results obtained from this study are encouraging to perform computationally intensive jobs on a distributed parallel architecture. Future work in this regard could be to use high performance computing to develop a dynamically linked BMP selection and placement model that distributes the various processes during a particular generation into multiple slave processors and the outputs from the slave processors is received by the master processor that performs various optimization routines to the variables based on the objective function values and therefore considerably reducing the computation time.
REFERENCES


Bekele, E. G., and J. W. Nicklow (2005), Multiobjective management of ecosystem services by integrative watershed modeling and evolutionary algorithms, Water Resources Research, 41(10), 10406.


Maringanti, C., I. Chaubey, and J. Popp (2009), Development of a Multi-Objective Optimization Tool for the Selection and Placement of BMPs for Nonpoint Source Pollution Control, Water Resources Research, (Accepted).


Table 1. Computation time on Windows and Unix machines for various SWAT model simulations used in the study

<table>
<thead>
<tr>
<th>Watershed</th>
<th>Years of simulation</th>
<th>No. of HRUs</th>
<th>Computation time</th>
<th>Windows</th>
<th>Unix</th>
</tr>
</thead>
<tbody>
<tr>
<td>L’Anguille River</td>
<td>18</td>
<td>433</td>
<td></td>
<td>10 mins</td>
<td>30 secs</td>
</tr>
<tr>
<td>Wildcat Creek</td>
<td>7</td>
<td>403</td>
<td></td>
<td>4 mins</td>
<td>15 secs</td>
</tr>
<tr>
<td>Lincoln Lake</td>
<td>28</td>
<td>1461</td>
<td></td>
<td>30 mins</td>
<td>3 mins</td>
</tr>
</tbody>
</table>
Figure 1. Pareto-optimum front progress during multi-objective optimization
Figure 2. Land use distribution of Lincoln Lake Watershed (Source: Chiang et al, 2008).
Figure Error! No text of specified style in document. Location of L’Anguille River Watershed in Arkansas and the location of USGS stream gauging stations in the watershed.
Figure Error! No text of specified style in document. Location of Wildcat Creek Watershed in Indiana and the observed gauge locations.
Figure 5. Flow chart that shows the methodology used in the study to distribute the computationally intensive jobs (dashed arrows) to multiple processors.
Figure 6. Change in the objective function as the number of generation increases for 5 different annual averages considered for sediment in LLW.
Figure 7. Progress of the Pareto-optimal front during optimization of the model for pesticide control in LRW.
Figure 8. Performance of the solutions obtained from the BMP tool when implemented at the watershed level using the SWAT model to simulate the pesticide loads.
Figure 9. Pareto-optimal fronts during the optimization (Left, the numbers indicate the generation number for the provided spread) and after the final generation (Right) for Sediment, Phosphorus, and Nitrogen reduction
Figure 10. Simulation of the solutions obtained during optimization using the SWAT model for estimating the (a) sediment, (b) phosphorus, and (c) nitrogen loads at Palestine gage station in the watershed.