GPU-based high-performance computing for integrated surface - sub-surface flow modeling

Phong V. V. Le[a], Praveen Kumar[a,b,*], Albert J. Valocchi[a], Hoang-Vu Dang[c]

[a]*Department of Civil and Environmental Engineering, University of Illinois, Urbana, IL 61801, USA*
[b]*Department of Atmospheric Sciences, University of Illinois, Urbana, IL 61801, USA*
[c]*Department of Computer Science, University of Illinois, Urbana, IL 61801, USA*

Abstract

The widespread availability of high-resolution lidar data provides an opportunity to capture micro-topographic control on the partitioning and transport of water for incorporation in coupled surface - sub-surface flow modeling. However, large-scale simulations of integrated flow at the lidar data resolution are computationally expensive due to the density of the computational grid and the iterative nature of the algorithms for solving nonlinearity. Here we present a distributed physically based integrated flow model that couples two-dimensional overland flow and three-dimensional variably saturated sub-surface flow on a GPU-based (Graphic Processing Unit) parallel computing architecture. Alternating Direction Implicit (ADI) scheme modified for GPU structure is used for numerical solutions in both models. Boundary condition switching approach is applied to partition potential water fluxes into actual fluxes for the coupling between surface and sub-surface models. The algorithms are verified using five benchmark problems that have been widely adopted in literature. This is followed by a large-scale simulation using lidar data. We demonstrate that the method is computationally efficient and produces physically consistent solutions. This computational efficiency suggests the feasibility of GPU computing for fully distributed, physics-based hydrologic models over large areas.

Keywords: Surface - sub-surface interactions, GPU computation, ADI scheme, finite difference, lidar

*Corresponding author
Email address: kumar1@illinois.edu (Praveen Kumar)

1. Introduction

The interaction between surface and sub-surface flow is an important component of the hydrologic cycle (Winter et al., 1998; Sophocleous, 2002). Capturing these interactions in models is thus critical to predicting soil moisture states and the responses of ecohydrological processes to global change across various scales (Rodriguez-Iturbe, 2000). Several conjunctive hydrologic models have been developed to integrate surface and sub-surface flow and are being used to address a range of science questions (Paniconi & Wood, 1993; Morita & Yen, 2002; Panday & Huyakorn, 2004; Ivanov et al., 2004; Kumar et al., 2009; Camporese et al., 2010; Shen & Phanikumar, 2010). These models have evolved into a family of coupling schemes that can represent the relevant physical processes influencing hydrologic responses from small catchment to larger river basin scales (Maxwell et al., 2014). In addition, these conjunctive models are being coupled to vegetation-hydrology dynamics (Ivanov et al., 2008), solute transport (VanderKwaak & Sudicky, 1999; Weill et al., 2011), and land-surface and atmospheric models (Maxwell & Miller, 2005; Maxwell et al., 2007). However, existing models have not been applied to capture the micro-topographic controls revealed by light detection and ranging (lidar) digital elevation model (lDEM) data (Le & Kumar, 2014). The goal of this paper is to present numerical scheme suited for Graphic Processing Unit (GPU) based computation to enable studies using lDEM over large areas.

The increasing availability of high-resolution topographic data from lidar technique has offered new opportunities for broader exploration of the control of landscape variability at fine scales such as water and nutrient dynamics (Lefsky et al., 2002; Schwarz, 2010; Ussyshkin & Theriault, 2011; Le & Kumar, 2014) and to explore behavioral response (Kumar, 2011). Previous studies have shown that depressions arising from micro-topographic variability can have significant effects on streamflow generation (Dunne et al., 1991; Frei et al., 2010; Thompson et al., 2010; Loos & Elsenbeer, 2011), soil moisture dynamics (Simmons et al., 2011), and the surface - sub-surface flow interactions (Frei & Fleckenstein, 2014). Recent work has begun to identify and characterize the spatial distribution of topographic depressions as power laws for size and volume, using lidar data (Le & Kumar, 2014). Dynamics associated with these micro-topographic features need to be incorporated into conjunctive surface - sub-surface flow models to understand their impacts on the hydrologic and biogeochemical processes. This incorporation also leads to a significant increase in computational cost for numerical models due to the size of the computational grid and the iterative nature of the algorithms in the coupled models.

A number of effort have contributed to the development of parallel formulation for some existing surface - sub-surface flow (Hwang et al., 2014; Kollet et al., 2010; Maxwell, 2013) and other coupled hydrologic systems (Gasper et al., 2014; Hammond et al., 2014). This has established the feasibility of high-resolution simulations at regional and continental scales. In addition, several studies have dealt with the computational scaling issues ranging from multi-cores to thousands of CPU cores on supercomputing systems (Gasper et al., 2014; Hammond et al., 2014; Kollet et al., 2010). This has also established the feasibility of high performance CPU computing for a range of applications for hydrologic modeling.

For the past few years, the graphics processing units (GPUs) have become increasingly popular and an integral part of today's mainstream computing systems (Owens et al., 2008). The increased capabilities and performance of recent GPU hardware in combination with high level GPU programming languages such as NVIDIA's Compute Unified Device Architecture (CUDA) and Open Computing Language (OpenCL) has provided massively parallel processing power for numerically intensive scientific applications, and made general purpose GPU computing accessible to computational scientists. It also opens a possibility for simulations over larger computational grids, for example, detailed ecohydrologic modeling over large domains at lidar-data resolution and large-scale computational fluid dynamics (Vanka, 2013). In comparison with central processing units (CPUs), however, GPUs have a distinct architecture centered around a large number of fine-grained parallel processors (Kirk & Hwu, 2010). Therefore, numerical models must be specifically structured such that processes are executed concurrently across many fine-grained processors.

This study aims to present an integrated flow model that couples two-dimensional overland flow and three-dimensional variably saturated sub-surface flow on a GPU-based parallel computing architecture (GCS-flow). The goal is to support simulations over large areas using fine resolution lDEM to reveal flow and accumulation associated with microtopographic features. Because the programmable units of GPU follow a single-instruction multiple-data (SIMD) model, we use finite difference alternating direction implicit (ADI) approach for discretizing independent tridiagonal linear systems and efficiently solving the governing equations. Though ADI for multi-dimensional nonlinear problems has been rarely used in favor of fully implicit methods using Krylov-based solvers with preconditioning due to stability, its has advantages over Krylov solvers in terms of scalability for large domains and computational cost as tridiagonal linear systems can be solved directly. Since data parallelism in ADI is abundant, there is high potential for this scheme to be advantageous on the throughput-oriented design of GPU.

The rest of the paper is organized as follows. In section 2, we provide an overview of the theory and numerical formulation of the GCS-flow model using ADI for GPU computing structure. Benchmark tests for model verification against other published solutions are presented in section 3. Results and analyses for simulations using lidar topographic data are shown in section 4. We demonstrate that the implemented model in GPU enables much faster execution than single-threaded performance in CPU. In section 5, the paper closes with the summary and discussion of the key points.

2. Theory and numerical methods

The theory of coupled surface - sub-surface flow has been an important area of research in the field of hydrology. Overview and details of the literature may be found in previous work (Paniconi & Wood, 1993; Morita & Yen, 2002; Panday & Huyakorn, 2004; Camporese et al., 2010; Sulis et al., 2010; Maxwell et al., 2014). We only provide a brief summary of the governing equations that form the basis for the set of coupled equations, the numerical method structured specifically for GPU architecture, and the coupling strategy between surface and sub-surface domains.

2.1. Overland flow

Overland flow is described by the depth-averaged flow equations commonly referred to as St. Venant equations that consist of a continuity (mass conservation) and two momentum equations. The continuity equation is written as:

$$\frac{\partial h}{\partial t} + \nabla \cdot (\boldsymbol{\nu} h) + q_e + q_r = 0 \tag{1}$$

where h is water depth on the surface $[L]$, t is the time $[T]$, $\boldsymbol{\nu}$ is depth averaged velocity vector $[L\ T^{-1}]$, q_e represents exchange fluxes between surface and sub-surface domains $[L\ T^{-1}]$, and q_r is a general source/sink term $[L\ T^{-1}]$ such as precipitation and evaporation. In diffusion flow, the momentum equations for overland flow reduce to:

$$S_{f,i} = S_n \tag{2}$$

where $S_{f,i} = \partial H/\partial x_i$ are friction slopes [-], i stands for the x- and y-directions, and S_n is the slope of the water [-] computed as $\sqrt{(\partial H/\partial x)^2 + (\partial H/\partial y)^2}$. The term $H = h + z$ is the water above a datum [L] and z is ground elevation above a datum [L] (Lal, 1998a,b).

Hromadka & Lai (1985) showed that Manning's equation can be used to establish a flow depth-discharge relationship:

$$\nu_i = -\frac{h^{2/3}}{n_b\sqrt{S_n}}\frac{\partial H}{\partial x_i} = -\frac{D}{h}\frac{\partial H}{\partial x_i} \tag{3}$$

in which D is the diffusion coefficient [$L^2\ T^{-1}$] and described by:

$$D = \begin{cases} \frac{h^{5/3}}{n_b\sqrt{S_n}}, & \text{for} \quad |S_n| > S_{min} \quad \text{and} \quad h > h_{min} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where n_b is the Manning's coefficient [$L^{-1/3}\ T$]. The condition $h > h_{min}$ is used to facilitate wetting and drying, and S_{min} is used to maintain D within finite limits (Lal, 1998a). Using Equation (3), the governing continuity equation of the overland flow in two spatial dimensions can be expressed as:

$$\frac{\partial H}{\partial t} = \frac{\partial}{\partial x}\left(D\frac{\partial H}{\partial x}\right) + \frac{\partial}{\partial y}\left(D\frac{\partial H}{\partial y}\right) - q_e - q_r \tag{5}$$

where x and y are the horizontal coordinates [L]. The term D is useful in linearizing and simplifying the diffusion flow equations. A variety of numerical algorithms can be used to solve the linearized diffusion overland flow equation (Lal, 1998a).

2.2. Variably saturated groundwater flow

The governing equation for variably saturated groundwater flow is represented on the basis of the mixed form Richards equation (Richards, 1931) as:

$$S_s\frac{\theta}{\phi}\frac{\partial \psi}{\partial t} + \frac{\partial \theta}{\partial t} = \nabla \cdot K(\theta)\left[\nabla\psi + \hat{\boldsymbol{k}}\right] + q_s + q_e \tag{6}$$

where ψ is the sub-surface pressure head [L], θ is the soil moisture [-], ϕ is the porosity [-], $\hat{\boldsymbol{k}}$ is the unit-upward vector, S_s is the specific storage coefficient [L^{-1}], K is unsaturated hydraulic conductivity [$L\ T^{-1}$], q_s is a general source/sink term representing pumping or injection [T^{-1}], and q_e represents the unit exchange fluxes between surface and sub-surface domains [T^{-1}] The ratio θ/ϕ is known as the degree of saturation.

The mixed form of the variably saturated flow equation has been shown to possess conservation property to maintain mass balance (Allen & Murphy, 1985; Celia et al., 1990). Different numerical methods can be used for solving variably saturated groundwater flow (Huyakorn & Pinder, 1983). In the mixed form Richards formulation presented here, a closed-form model by van Genuchten (1980) is used to describe the constitutive relationships between θ, ψ, and K. The water retention curve is given by:

$$\Theta = \frac{\theta(\psi) - \theta_r}{\theta_s - \theta_r} = \left[\frac{1}{1 + (\alpha\psi)^n}\right]^{1-1/n} \tag{7}$$

where Θ is the relative saturation [-], θ_r is residual water content [-], θ_s is saturated water content [-] (often approximated by the porosity ϕ), n is the pore-size distribution [-], and α is a parameter related to the inverse of the air entry suction $[L^{-1}]$. The unsaturated hydraulic conductivity function is given by (Mualem, 1976):

$$K(\theta) = K_s \Theta^{\frac{1}{2}} \left[1 - \left(1 - \Theta^{1-\frac{1}{n}}\right)^{1-\frac{1}{n}}\right]^2 \tag{8}$$

where K_s is the saturated hydraulic conductivity $[L\,T^{-1}]$ identified from soil physical properties.

2.3. Discretization and numerical implementation

The alternating direction implicit (ADI) method is used for numerical solutions in both surface and sub-surface flow models in GCS-flow. This approach has advantages over the fully implicit methods in terms of simplicity and cost (on a per iteration basis) because only tridiagonal linear systems are required to provide direct solutions. In addition, the discretization of ADI is more scalable ($O[\mathcal{N}]$) than fully implicit approach ($O[\mathcal{N}^d]$) as the problem dimensions increase, in which $\mathcal{N}$ and d represent the size and the number of dimensions of the domain. An et al. (2011) have compared the performances between ADI and preconditioned conjugated gradient methods for multi-dimensional variably saturated flow implemented on CPU. They showed that ADI method is faster than fully implicit method while still yielding very similar results. However, the main disadvantage of ADI is the constrains in stability which requires smaller time steps than unconditionally stable fully implicit methods. ADI has better stability condition than explicit method without hard requirements on the time step. Morita & Yen (2002) showed the stability criterion of ADI for 2D overland flow, $\xi_1 = D\Delta t \left(\Delta x^{-2} + \Delta y^{-2}\right) < 5$, and 3D subsurface flow, $\xi_2 = K\Delta t \left(\Delta x^{-2} + \Delta y^{-2} + \Delta z^{-2}\right) < 1.25$, where Δx, Δy, Δz are the grid spacing in their respective directions and D and K are shown in Equations (5) and (6), respectively.

The mass balance condition with Crank-Nicolson type scheme forms the basis for the ADI formulation in overland flow. Using the ADI method, Lal (1998a) showed that the continuity equation (1) for overland flow can be expressed in the following split formulation in sequence:

$$(1-\delta_x)H_{i,j}^{n+\frac{1}{2}} = (1+\delta_y)H_{i,j}^{n} - \frac{\Delta t}{2}(q_e + q_r) \tag{9a}$$

$$(1-\delta_y)H_{i,j}^{n+1} = (1+\delta_x)H_{i,j}^{n+\frac{1}{2}} - \frac{\Delta t}{2}(q_e + q_r) \tag{9b}$$

where n is the time step, (i, j) denotes spatial location, δ_x and δ_y are the standard second-order centered differencing operators in x and y direction, respectively. In our model, the coupled Equations (9) are solved as two 1-D problems for each row and column of the 2-D domain for tridiagonal matrices at every half time step $\frac{\Delta t}{2}$. Linearized implicit methods use D values of the previous time step (Lal, 1998a). Right-hand sides of these equation consist of entirely known values at the time of computation. The detailed derivation and numerical form of Equations (9) are presented in the Appendix A.

A simple mass-conservative numerical approach based on the backward Euler scheme associated with Picard iteration (Celia et al., 1990) is modified for 3-D sub-surface flow using the ADI method in this study. Because the relationship between θ and ψ is highly non-linear, iterative calculation and linearization are needed to solve these systems. The backward Euler approximation for 3-D variably saturated groundwater flow can be written as:

$$\frac{S_s}{\phi}\theta^{n+1,m}\left[\frac{\psi^{n+1,m+1}-\psi^{n}}{\Delta t}\right] + \left[\frac{\theta^{n+1,m+1}-\theta^{n}}{\Delta t}\right] = \frac{\partial}{\partial x_i}\left(K^{n+1,m}\,\frac{\partial \psi^{n+1,m+1}}{\partial x_i}\right) - \frac{\partial K^{n+1,m}}{\partial x_3} + q_e + q_s = 0 \tag{10}$$

where m is the Picard iteration level. Values at level m are known while at level $m+1$ are unknown. Here x_i denotes spatial coordinates.

Using ADI, we sequentially solve Equation (10) at every $\frac{\Delta t}{3}$ time step, keeping one direction implicit and the other two explicit. The implicit direction is then changed to the next direction (or axis), and so on until the next time step. The derivation and full numerical form of Equations (10) separated in x, y, and z direction using ADI are presented in Appendix B. The iteration process to solve Equations (10) continues until the difference between the calculated values of the pressure head of two successive iteration levels become less than the user-defined tolerance for convergence:

$$\left|\psi^{n+1,m+1} - \psi^{n+1,m}\right| \leq \epsilon_\psi \tag{11}$$

Independent linear systems obtained from ADI for the two models are suitable for parallelizing in a large number of fine-grained processors in GPU devices.

2.4. GPU Parallelization

We implement the integrated surface - sub-surface flow model on a GPU parallel computing structure. The model supports the use of different generations and types of CUDA-capable GPUs, which consists of a sequential *host* program that may execute parallel programs known as *kernels* on a parallel *device*. While data processing is performed on the host using C++ programming language, all model computation is executed in parallel on NVIDIA's GPUs (device) using CUDA programming language (Nickolls et al., 2008). CUDA virtualizes multiprocessors as blocks and processors as threads, which enables users to run a potentially large number of parallel threads and blocks across different generations of GPUs regardless of the number of physical processors (Zhang et al., 2010). Each thread runs the same scalar sequential program for solving tridiagonal linear systems (ADI solvers).

We sequentially solve the 2-D overland flow and 3-D sub-surface flow sub-models in parallel using these ADI solvers and couple them through an iterative strategy presented in the next section. More specifically, for each model we set up and solve simultaneously a large number of systems of n linear equations of the form $Ax = b$, where A is the tridiagonal matrix, and x and b are vectors. This approach discretizes each governing equation for both sub-models into a number of independent tridiagonal linear systems which can be solved simultaneously using parallel cyclic reduction (PCR) method (Hockney & Jesshope, 1988) and the Thomas (TDMA) algorithm (Thomas, 1949). To efficiently solve these systems in parallel, we map the PCR solvers to the GPU's two-level hierarchical architecture with systems mapped to blocks and equations mapped to threads to utilize shared memory. If matrix size is small enough (i.e. in vertical z direction), TDMA solvers are mapped directly to threads to utilize local and register memory (Figure 1). The number of systems for large simulation domain we solve is usually far larger than the number of multiprocessors, so that all multiprocessors are fully utilized.

At the thread level, the total storage consists of five main vectors: three for the matrix diagonals, one for the right-hand side, and one for the solution vector. These five vectors store the data of all systems continuously, with the data of the first system stored at the beginning of the arrays, followed by the second system, the third system, and so on. For each system, we load the three diagonals and right-hand side from global memory to register, local, or shared memories (Figure 1), solve and store the solution back to global memory on device. Therefore, global memory communication only occurs at the beginning and end of each time step in ADI solver. Other vectors (i.e. for the linearization of diffusion coefficients and van Genuchten relationship) can be generated in threads for the solutions of surface and sub-surface flow. As data is stored in global memory, no communication between CPU and GPU is needed as the direction (or axis) of calculation is changed.

2.5. Coupled surface - sub-surface formulation

A boundary condition switching procedure (Paniconi & Wood, 1993; Camporese et al., 2010; Sulis et al., 2010; Camporese et al., 2014) is used for coupling the surface and sub-surface flow in GCS-flow model. Specifically, the boundary condition at any surface ground nodes of the sub-surface domain is allowed to switch between a Dirichlet and a Neumann type, depending on the saturation (or pressure head) state of that node. A Neumann (or specified flux) boundary condition corresponds to atmosphere-controlled infiltration or exfiltration with the flux equal to the rainfall or potential evaporation rate given by the atmospheric input data. In contrast, a Dirichlet (specified head) boundary condition is activated when the surface node reaches a threshold level of saturation (and ponding) or lower moisture deficit and the infiltration and exfiltration processes become soil-limited. This switching algorithm is applied for both rainfall and evaporation cases. We refer to previous studies (Camporese et al., 2010, 2014) for further details on boundary condition switching approach.

Flows between the sub-surface and overland flow domains are represented through the unit interactive flux q_e across ground surface in Equations (1) and (6). Through this term, the coupling strategy we used partitions potential (atmospheric) fluxes into actual fluxes across the land surface and changes in surface storage. In the surface model, the surface to sub-surface contribution and water depth are determined after solving the overland flow equation for subsequent input to the sub-surface flow equation, while the sub-surface to surface contribution is determined after solving the sub-surface flow equation for subsequent procedure for the solution of the coupled equation. The atmospheric fluxes are resolved only in subsurface model. The exchange of flux performed via the switching algorithm in the sub-surface module and the simple mass balance calculation in the surface module resolves the coupling in the model without the need to introduce new parameters representing an exchange process or an interface property but still guarantees the necessary continuity of flux and pressure head at the ground surface (Camporese et al., 2010).

3. GPU computing performance

To demonstrate the efficiency of GPU computing for ADI approach, we compared the performance of iterative ADI solvers in 5 simulations for unsaturated sub-surface flow between a single-threaded sedec-core CPUs (16 Intel Xeon 2.67 GHz processors, written in C++) and each of two NVIDA GPUs (Tesla C2070 and Tesla K40, written in CUDA C++). The simulation domains are isotropic, rectangular, and set at different dimensions ($N_x \times M_y \times P_z$): (i) $78 \times 78 \times 10$; (ii) $128 \times 128 \times 16$; (iii) $256 \times 256 \times 16$; (iv) $512 \times 512 \times 16$, and (v) $1024 \times 1024 \times 16$, where N_x, M_y, and P_z are the numbers of soil layers or grids in x, y, and z directions, respectively. The mesh discretization and time stepping are identical for all 5 cases: $\Delta x = 5$ $[m]$, $\Delta y = 5$ $[m]$, $\Delta z = 0.2$ $[m]$, and $\Delta t = 0.05$ $[hr]$. Simulations are run for 48 $[hr]$. Parameters for the closed-form equation for the soil water retention curve and unsaturated hydraulic conductivity function are obtained from previous study (Celia et al., 1990), where $\alpha = 0.0335$ [cm^{-1}], $\theta_s = 0.368$, $\theta_r = 0.102$, $n = 2$, $K = 0.0332$ [m/hr]. Initial and boundary conditions were taken as $\psi(x, y, z, 0) = -5.0$ [m], $\psi(x, y, P_z, t) = \psi_{bottom} = -3.0$ [m], $\psi(x, y, 0, t) = \psi_{top} = -1.0$ [m], no-flow boundaries are used for horizontal flow, and no source and sink terms are included. While C2070 devices (second-generation) have 448 CUDA cores and deliver up to 515 gigaflops of double-precision peak performance, K40 devices (third-generation) are configured with 2,880 CUDA cores and deliver up to 1,430 gigaflops of double-precision peak performance (NVIDIA Corporation, 2011).

Figure 2 shows the relative speed-up for solving ADI using tridiagonal matrix systems in two GPU devices over that in CPU. The average speed-up of the simulations for C2070 and K40 are 26.3 and 83.2, respectively. The speedup comes from the ability of single instruction, multiple thread architecture (Kirk & Hwu, 2010) in GPUs to simultaneously execute thousands of linear systems solver. However, since each GPU core is clocked at as low as few hundreds Mhz, and latency due to fetching for matrix entries is limited by the memory subsystems, the speedup is not close to the number of GPU cores. Nevertheless, we achieve a performance typically seen in GPU computing (Lee et al., 2010). We also see a large difference between the two GPU generations in this comparison. The K40 device with improved architecture to accelerate computation, higher theoretical peak-performance and number of processors is 3 to 6 times faster than the C2070 device. The discrepancy in performance is also found among the size of the computational grid. Larger domains of simulation tend to get better speed-up than smaller domains as the occupancy of GPUs processor is higher. Our scalable ADI solver implemented in this work is also expected to utilize the architecture improvement of next CUDA-capable GPU generations.

4. Benchmark simulation tests

Due to the lack of an analytical solution for coupled surface - sub-surface flow, we use a set of benchmark test cases summarized below to compare our model with those published by others for verification. We use direct value comparison method (Bennett et al., 2013) for measuring the quantitative performance among models in all test cases. Detailed information about these tests can be found in previous work (Gottardi & Venutelli, 1993; Panday & Huyakorn, 2004; Kollet & Maxwell, 2006; Kumar et al., 2009; Sulis et al., 2010; Maxwell et al., 2014). The test cases include: (i) tilted V-catchment, (ii) infiltration excess, (iii) saturation excess, (iv) slab tests, and (v) return flow. These involve simple geometries associated with other features (topography, hydraulic and hydrogeological properties, and atmospheric forcing), but with complex physical responses designed to thoroughly compare model behavior (Maxwell et al., 2014). The test cases also feature step functions of rainfall followed by a recession period. The response variables analyzed include domain outflow, saturation conditions, and location of intersection between the water table and land surface (Maxwell et al., 2014). Model parameters used for these tests are similar to the set shown in two inter-comparison studies by Maxwell et al. (2014) and Sulis et al. (2010) and presented in Table 1. To avoid confusion, we only select four representative models based on similarities and differences for comparison in this paper: ParFlow (Parallel Flow) - uses fully implicit finite difference method for numerical solution (Kollet & Maxwell, 2006); Cathy (Catchment HYdrology) - uses finite element method and the boundary condition switching approach for coupling strategy (Paniconi & Wood, 1993; Paniconi & Putti, 1994); tRIBS+VEGGIE (Triangulated Irregular Network (TIN)-Based Real Time Basin Simulator) - uses an irregular spatial discretization and first-order exchange for coupling strategy (Ivanov et al., 2004); and PAWS (Process-based Adaptive Watershed Simulator) - uses asynchronous linking and couples 1-D Richard equation in unsaturated zone with saturated domain (Shen & Phanikumar, 2010). Additional comparisons for similar tests from other available models can be found in previous studies (Maxwell et al., 2014; Qu & Duffy, 2007).

4.1. Tilted V-catchment

The tilted V-catchment problem is a standard test case for the overland flow model. The domain consists of two inclined planar rectangles of width 800 $[m]$ and length 1000 $[m]$ connected together by a 20 $[m]$ wide sloping channel as shown in Figure 3a & b. This test only considers the surface flow processes and is used to assess the behavior of the surface routing component without any contribution from the sub-surface by assuming that no infiltration occurs. The slope of the planes is 5% and the slope of the channel is 2%. The simulation consists of a 90 $[min]$ rainfall event (at a uniform intensity of 1.8×10^{-4} $[m/min]$) followed by 90 $[min]$ of recession (Figure 3c).

The comparisons of outflow result from GCS-flow with other four models in the tilted V-catchment test case are shown in Figure 3d. The GCS-flow model generally predicts quite similar behaviors to the four models selected. GCS-flow exhibits agreement with tRIBS-VEGGIE for rising limb, prediction of time to steady state, peak flow, and recession phases. However, the largest model differences during the rising phase are found in the predictions of GCS-flow and Parflow model. We have also found that outflow in GCS-flow occurs slightly earlier than all other models during the rising limb phase of the hydrograph. This discrepancy may be attributed to the time-splitting treatment of ADI for diffusive flow in comparison with other overland routing models. Nonetheless, there is a greater agreement among all models during the recession phase.

4.2. Infiltration Excess

The infiltration excess tests aim to investigate the Hortonian runoff produced before complete saturation of the soil column. This is achieved by specifying homogeneous saturated hydraulic conductivity K_s smaller than the rainfall rate. We test the model with two different low values of K_s as shown in Table 1. The domain is an inclined planar rectangle of width 400 $[m]$ and length 320 $[m]$ (Figure 4a) with a uniform soil depth of 5 $[m]$. The slope of the planes in $x-$direction and $y-$direction are 0.05% and 0%, respectively. No-flow boundary is prescribed at bottom of the domain, and the initial water table is set at 1 $[m]$ depth. A rainfall event 200 $[min]$ in duration with a rate of 3.3×10^{-4} $[m/min]$ was applied to generate runoff, followed by 100 $[min]$ of recession period (Figure 4b). Model outflow is measured at the outlet of the grey strip of cells (Maxwell et al., 2014; Sulis et al., 2010).

Figure 4c shows the outflow rate of GCS-flow as a function of time from the infiltration excess test as compared with other models. In general, the four models produce very similar hydrograph behavior throughout all phases as well as the magnitude of the outflow for both values of K_s. For the lower K_s test case, the largest difference in outflow is found again between GCS-flow and Parflow during the rising limb phase. We observe that the outflow is larger than Parflow at the beginning but the two models tend to converge at the end of the rising limb. This discrepancy is similar to the overland V-catchment test case shown above. For the higher K_s test case, in which overland flow is less dominant, the outflow obtained from GCS-flow is in better agreement with the four models. However, we also see that the recession curve drops slightly faster than other models and most discrepancy is with Parflow in the last 60 $[min]$ of the simulation time. Associated with overland flow, the sharper drop of the recession curve may be attributed to the infiltration that is treated using ADI and boundary switching algorithm in GCS-flow model.

4.3. Saturation Excess

The saturation excess tests are designed to investigate the Dunne runoff produced by ensuring complete saturation of the soil column from below and the intersection of the water table with the land surface. This is also achieved by specifying a homogeneous saturated hydraulic conductivity ($K_{sat} = 6.94 \times 10^{-4}$ $[m/min]$) which is larger than the rainfall rate (Table 1). Boundary conditions and domain of simulation are the same as the infiltration excess test (Figure 4a). However, we run the model for two different values of initial water table depth at: 0.5 $[m]$ and 1.0 $[m]$ as shown in Table 1. The test case with water table depth near the ground is expected to produce runoff earlier and will be associated with larger flow magnitude than the test with deeper water table level. As in the previous test, a rainfall event 200 $[min]$ in duration with a rate of 3.3×10^{-4} $[m/min]$ was applied to generate runoff, followed by 100 $[min]$ of recession Figure 4b.

Figure 4d shows the outflow rate of GCS-flow as a function of time from the saturation excess test and how it compares with other four models. We observe that the hydrograph produced from GCS-flow model is in most agreement with tRIBS+VEGGIE and PAWS for both tests. In addition, difference between GCS-flow and Cathy and Parflow is smaller than in the infiltration excess test, especially for shallow water table case. Outflow occurs at very similar time for both shallow water table tests ($\sim$20 $[min]$) and deep water table test ($\sim$120 $[min]$). Peak flow is also found in good agreement with all models in the shallow test. For deep water table test case, GCS-flow is in largest disagreement with Cathy during the rising limb. The difference in peak flow between GCS-flow and Cathy is about 1.1 [m^3/min] (12%). This is quite surprising as the two models use the same boundary switching approach for calculating infiltration. But we note that Cathy has the lowest peak flow among all the models. The discrepancy may comes from the numerical method used for both overland and subsurface flow in the models that need further investigations.

4.4. Slab Test

The slab test illustrates the effect of spatial heterogeneity of soil hydraulic conductivity in the same domain as in the infiltration and saturation excess tests. In this simulation, the soil is generally uniform (with a K_s value of 6.94×10^{-4} [m/min]) except for a 100-m long, 0.05-m thick, very low conductivity slab with $K_s = 6.94 \times 10^{-6}$ [m/min] as shown in Figure 5a. The saturated hydraulic conductivity of the slab is designed to generate infiltration excess runoff while the hydraulic conductivity of the rest of the domain is large and will only generate surface runoff through saturation excess. Boundary conditions and domain of simulation are the same as the previous two tests, and water table is set at 1 [m] depth. As in the infiltration excess case, a rainfall event 200 [min] in duration with a rate of 3.3×10^{-4} [m/min] was applied to generate runoff, followed by 100 [min] of recession (Figure 5b). We expect the combination of both infiltration excess and saturation excess runoff in outflow for this test.

The comparisons of outflow result from GCS-flow with other models in the slab test case are shown in Figure 5c. We found differences between the GCS-flow model and all others in this test. First, runoff occurs after 80 [min] which is later than tRIBS+VEGGIE model but earlier than PAWS, Parflow, and Cathy models. Second, During the rising limb phase, the hydrograph curve from GCS-flow is quite smooth while ones produced from other models are quite flat. Finally, the peak flow from GCS-flow model is closer to Cathy model (1.1 [m^3/min]) and lower than the other two models. Both GCS-flow and Cathy models are similar in using the boundary condition switching approach, which might explain this similarity in the response of runoff. During the recession phase, similar to other tests, outflow in GCS-flow model is in much better agreement with other models.

Snapshots of saturation profile obtained from GCS-flow model at: 0, 60, 90, and 150 [min] are presented in Figure 6. These moments of time, before the recession period, are chosen to show the complex physical responses of heterogeneous soil columns to infiltration, saturation, and lateral unsaturated flow. While water table in soil columns with uniform K_s rises quickly due to saturation excess, water table in the soil columns with heterogenous K_s (slab on top) rises very slowly due to infiltration capacity limits. Lateral unsaturated flow is also observed due to gradient of moisture in the sub-surface. Saturation profile for other models can be found in Maxwell et al. (2014, see Figure 8). We also found some differences in saturation profiles among all models. These may likely be explained due to the different coupling strategies and numerical scheme for solving the models.

4.5. Return flow

This test case uses the same hillslope domain as for the infiltration and saturation excess tests (see Figure 4a) but with much higher values of hydraulic conductivity ($6.94 \times 10^{-2}[m/min]$) to allow rapid rise and fall of the water table. The water table is initially set at 0.5 $[m]$ depth. Return flow is generated by an atmospheric forcing sequence formed by an initial 200 $[min]$ rainfall event of uniform intensity $1.5 \times 10^{-4}[m/min]$ followed by 200 $[min]$ of evaporation at a uniform rate of $5.4 \times 10^{-6}[m/min]$ (Figure 7a). Two hillslope inclinations are considered (0.5% and 5%) to highlight the effect of different characteristic times scales of the surface and subsurface processes (Maxwell et al., 2014; Sulis et al., 2010). The dynamics of return flow are evaluated by tracking the evolution of the intersection point between water table and the land surface. The model is also run with a uniform discretization comprising 100 vertical layers as done in other studies

Figure 7b&c show the intersection point between the water table and the land surface as a function of time for hillslope inclination of 0.5% and 5%, respectively. For the 0.5% slope case, in which infiltration and subsurface flow remain predominant, although the prediction of time for steady state is similar for all models, we observe that water table along the hillslope obtained from GCS-flow rises and intersects the ground more uniformly than other models. During evaporation phase, the recession of the water table is slower than other models, except for Cathy (sheet flow). For the 5% inclination case, in which the catchment drains faster, we do see more disagreements among GCS-flow and others during all rising, quasi-steady equilibirum, and recession phases. Specifically, the intersection point in GCS-flow are closer to the upslope point during steady state than others which implies that the moisture gradient resulting from the surface slope was not captured well in GCS-flow. This could be due to the discretization of finite difference in ADI for solving the variably saturated equation. This issue may also explain the uniform rising of water table before reaching steady state in both cases. The recession of the water table is faster than the gentler-slope case and exhibits better agreement with PAWS and Parflow.

In general, the GCS-model performs quite similarly to other conjunctive models that have been published in literature. Although, we observe greater differences among models in more complex tests and owing to different numerical approach, all model are very consistent in more simple tests. These results support the rationale for the modeling scheme developed using GPUs for larger scale simulation presented in the next section.

5. Simulations with lidar data

We run the GCS-flow model for an observed topography in the Goose Creek watershed of the Sangamon River Basin, in central Illinois, USA (Figure 8). This watershed is intensively managed for agriculture and is part of the Critical Zone Observatory for Intesively Managed Landscapes (http://imlczo.org). Lidar data used is available from the Illinois State Geological Survey (https://www.isgs.illinois.edu). The domain of simulation is 6.6 $km \times 7.4$ km with 2.0 m soil depth. Topographic resolution of the simulation domain is 1.2 $m \times 1.2$ m (Figure 8c). This results in over 35 million grid points on the surface and 350 million grid points over the entire subsurface domain. At this high resolution, micro-topographic features can be observed on the land surface such as road-side ditches, small depressions, etc. Topographic gradient in the study site is quite small as elevation variation is very small (from 205 to 222 $[m]$ above sea level, average slope $\approx 0.25\%$). Soil physical properties are available from Soil Survey Geographic database (SSURGO) distributed by Natural Resources Conservation Service (http://websoilsurvey.sc.egov.usda.gov). The average values of soil properties used in our simulation is follow: $K_s = 0.0054$ $[m/min]$, $\theta_s = 0.37$, $\theta_r = 0.10$, $\alpha = 3.35$ $[cm^{-1}]$, and $n = 1.25$.

Observed atmospheric forcing data obtained from nearby Bondville flux tower is used to drive the GCS-flow model. We use precipitation collected at 30 min intervals during a three week period in May 2005 for running the simulations (Figure 9a). Evaporation rate is assumed at constant rate 1 $[mm/day]$ for the domain. Initial soil moisture is set uniformly at 0.26 over the entire domain. In surface domain, no-flow boundaries are applied at the lateral boundaries. In sub-surface domain, we use a fixed boundary pressure head at -4 $[m]$ for the bottom and no-flow boundaries for the lateral interfaces. In term of computational efficiency, the simulation domain results in approximately 3.5

$\times 10^8$ unknowns (grid cells) for computation. The model takes about 19.6 [hr] on Tesla K40 GPUs for completion. Given using a single GPU device, this computational efficiency is significant and makes hydrologic simulations feasible over large areas.

A smaller area (900 $m \times 1080$ m) within the simulation domain is chosen (red box in Figure 8c) to show for detailed illustration of the vertical variation of soil-moisture as impacted by micro-topographic features. The snapshot of overland spatial flow in this reduced area after 320 [h] is presented in Figure 9b. The model shows flow accumulation in topographic depressions and in the road-side ditches which carries significant flows in these agricultural landscapes. These features are often ignored in modeling with lower resolution and coarser computational grid. The profile of soil moisture over depth at the same time is shown in Figure 9c. While the top boundary of the domain is controlled by surface water availability and atmospheric fluxes, the low values of bottom boundary allow water to drain significantly to soil layers below the simulation domain. We however observe the positive correlation between micro-topographic depressions on the land surface and soil moisture distribution below-ground. Areas where water is accumulated due to low elevation on the landscape provide more moisture for infiltration than other area.

6. Conclusion

A formulation of coupled surface - sub-surface flow model using lidar-resolution topographic data that is implemented on GPU parallel computing architecture has been presented. The numerical solution for both overland flow and sub-surface flow model is based on the alternating direction implicit (ADI) method. While 2-D classic ADI is applied for overland flow model, an iterative ADI associated with Picard iteration approach is used for 3-D sub-surface model due to the non-linearity in the relationship in the mixed form of Richards equation. This approach combines benefits in simplicity and cost for computation because only tridiagonal linear systems are involved for providing direct solution and the ability to decompose into fine-grain tasks for GPU parallel structure. The model has been compared with others using several standard benchmark test cases. The results from benchmark tests generally show good agreements among all the model for a wide variety of benchmark test cases. Some model differences are found in complex tests due to different coupling strategies and numerical solution.

The GCS-flow model has been used to simulate an intensively managed landscape in the Goose

Creek watershed, Illinois, USA. The lidar-derived topographic data at 1.2 m resolution is used for detailed hydrologic simulation. Results presented indicate that this performance is faster than CPU and has the potential to apply for detailed ecohydrologic modeling in large areas.

We suggest that future work should aim to expand this model to understand the dynamics and linkages between soil moisture and microtopographic features on a range of applications including ecohydrology, agriculture, etc. In addition, with rapid advances in GPU computing, the model can be used as a starting point to explore: (i) possible alternative formulations (i.e. fully implicit scheme, iterative methods) based improved computational libraries for GPU developed by the community; (ii) new memory structures and capabilities released in the next GPU generations for parallel computing in general and for hydrologic modeling in particular; and (iii) implementations and scaling behaviors of hydrologic and integrated flow modeling on multiple-GPUs. Such efforts will lead to an improved understanding and a more robust generation of integrated surface-subsurface flow modeling using high-performance GPU computing.

Appendix A: ADI discretization for 2D overland flow

The mass balance condition with Crank-Nicolson type scheme forms the basis for the ADI formulation in overland flow model. Following a study by Lal (1998a), the discretization of overland flow continuity equation may be written as:

$$H_{i,j}^{n+1} = H_{i,j}^{n} + \Delta t \left[\alpha \frac{Q_{net}(H^{n+1})}{\Delta x \Delta y} + (1-\alpha)\frac{Q_{net}(H^{n})}{\Delta x \Delta y} - q_e - q_r\right] \tag{A1}$$

in which α is the weighting factor for numerical scheme ($\alpha = 0.5$ for Crank-Nicolson), n is the time step, (i,j) denotes spatial location, and Q_{net} $[L^3]$ is the net inflow to the cell as a function of potential head, computed as:

$$\begin{aligned} Q_{net} = {} & D_{i+\frac{1}{2},j}(H_{i+1,j} - H_{i,j}) + D_{i-\frac{1}{2},j}(H_{i-1,j} - H_{i,j}) \\ & + D_{i,j+\frac{1}{2}}(H_{i,j+1} - H_{i,j}) + D_{i,j-\frac{1}{2}}(H_{i,j-1} - H_{i,j}) \end{aligned} \tag{A2}$$

where D is diffusion coefficient $[L^2\,T^{-1}]$ (See equation (4) in the main text). The spatial differencing operators used in the derivation are as follows:

$$\delta_x H_{i,j}^n = 0.5\frac{\Delta t}{\Delta x \Delta y}\left[D_{i+\frac{1}{2},j}(H_{i+1,j}^n - H_{i,j}^n) + D_{i-\frac{1}{2},j}(H_{i-1,j}^n - H_{i,j}^n)\right] \tag{A3a}$$

$$\delta_y H_{i,j}^n = 0.5\frac{\Delta t}{\Delta x \Delta y}\left[D_{i,j+\frac{1}{2}}(H_{i,j+1}^n - H_{i,j}^n) + D_{i,j-\frac{1}{2}}(H_{i,j-1}^n - H_{i,j}^n)\right] \tag{A3b}$$

After rearranging the unknown values to the left-hand side, Equations (A1) and (A2) can be now expressed using the standard second-order centered differencing operators as:

$$(1 - \delta_x - \delta_y)H_{i,j}^{n+1} = (1 + \delta_x + \delta_y)H_{i,j}^{n} - (q_e + q_r)\Delta t \tag{A4}$$

By neglecting higher-order terms, Equation (A4) can also be split into sequences:

$$(1 - \delta_x)H_{i,j}^{n+\frac{1}{2}} = (1 + \delta_y)H_{i,j}^{n} - \frac{\Delta t}{2}(q_e + q_r) \tag{A5a}$$

$$(1 - \delta_y)H_{i,j}^{n+1} = (1 + \delta_x)H_{i,j}^{n+\frac{1}{2}} - \frac{\Delta t}{2}(q_e + q_r) \tag{A5b}$$

The ADI finite-difference expressions for the overland flow can be written as:

$$\begin{aligned} &H_{i,j}^{n+\frac{1}{2}} - 0.5\frac{\Delta t/2}{\Delta x \Delta y}\left[D_{i+\frac{1}{2},j}\left(H_{i+1,j}^{n+\frac{1}{2}} - H_{i,j}^{n+\frac{1}{2}}\right) + D_{i-\frac{1}{2},j}\left(H_{i-1,j}^{n+\frac{1}{2}} - H_{i,j}^{n+\frac{1}{2}}\right)\right] = \\ &\qquad H_{i,j}^{n} + 0.5\frac{\Delta t/2}{\Delta x \Delta y}\left[D_{i,j+\frac{1}{2}}\left(H_{i,j+1} - H_{i,j}\right) + D_{i,j-\frac{1}{2}}\left(H_{i,j-1} - H_{i,j}\right)\right] + (q_e + q_r)\frac{\Delta t}{2} \end{aligned} \tag{A6}$$

$$\begin{aligned} &H_{i,j}^{n+1} - 0.5\frac{\Delta t/2}{\Delta x \Delta y}\left[D_{i,j+\frac{1}{2}}\left(H_{i,j+1}^{n+1} - H_{i,j}^{n+1}\right) + D_{i,j-\frac{1}{2}}\left(H_{i,j-1}^{n+1} - H_{i,j}^{n+1}\right)\right] = \\ &H_{i,j}^{n+\frac{1}{2}} + 0.5\frac{\Delta t/2}{\Delta x \Delta y}\left[D_{i+\frac{1}{2},j}\left(H_{i+1,j}^{n+\frac{1}{2}} - H_{i,j}^{n+\frac{1}{2}}\right) + D_{i-\frac{1}{2},j}\left(H_{i-1,j}^{n+\frac{1}{2}} - H_{i,j}^{n+\frac{1}{2}}\right)\right] + (q_e + q_r)\frac{\Delta t}{2} \end{aligned} \tag{A7}$$

The coupled Equations (A6) and (A7) are solved as two 1D problems for each row and column of the 2D domain using the TDMA or PCR algorithms for tridiagonal matrices at half time step $\frac{\Delta t}{2}$. Right-hand sides of these equation consist of entirely known values at the time of computation. The values of $H_{i,j}^{n+\frac{1}{2}}$ are obtained from Equation (A6) in the first-half time step and then used to solve Equation (A7) in the second-half time step.

Appendix B: ADI discretization for 3D variably saturated sub-surface flow

The backward Euler scheme associated with Picard iteration is one of the most widely used time approximation for the Richards equation and applied in this study. The two terms in the left-hand side of the variably saturated sub-surface flow equation are approximated as:

$$S_s\frac{\theta}{\phi}\frac{\partial \psi}{\partial t} + \frac{\partial \theta}{\partial t} \approx \frac{S_s}{\phi}\theta_{i,j,k}^{n+1,m}\left[\frac{\psi_{i,j,k}^{n+1,m+1} - \psi_{i,j,k}^{n}}{\Delta t}\right] + \left[\frac{\theta_{i,j,k}^{n+1,m+1} - \theta_{i,j,k}^{n}}{\Delta t}\right] \tag{B1}$$

Here, (i,j,k) denote spatial location in x, y, and z directions, respectively, n and m denote the time and the Picard iteration levels, respectively. After Celia et al. (1990), moisture content at new time step and a new iteration level $\theta_{i,j,k}^{n+1,m+1}$ is expanded using first-order truncated Taylor series, in terms of pressure-head perturbation, about the expansion point $\psi^{n+1,m}$ as follow:

$$\theta_{i,j,k}^{n+1,m+1} = \theta_{i,j,k}^{n+1,m} + \left.\frac{d\theta}{d\psi}\right|_{i,j,k}^{n+1,m} (\psi_{i,j,k}^{n+1,m+1} - \psi_{i,j,k}^{n+1,m}) + O(\delta^2) \tag{B2}$$

The specific water capacity function of the soil $C(\psi)$ $[L^{-1}]$ is defined as:

$$C(\psi) = \frac{d\theta}{d\psi} \tag{B3}$$

Using equation (B2) and (B3), the Equation (B1) can be expressed as:

$$S_s\frac{\theta}{\phi}\frac{\partial\psi}{\partial t} + \frac{\partial\theta}{\partial t} \approx \frac{S_s}{\phi}\theta_{i,j,k}^{n+1}\left[\frac{\psi_{i,j,k}^{n+1,m+1} - \psi_{i,j,k}^{n}}{\Delta t}\right] + \frac{\theta_{i,j,k}^{n+1,m} - \theta_{i,j,k}^{n}}{\Delta t} + C_{i,j,k}^{n+1,m}\left[\frac{\psi_{i,j,k}^{n+1,m+1} - \psi_{i,j,k}^{n+1,m}}{\Delta t}\right] \tag{B4}$$

Rearranging and use the increment in iteration: $\delta^m = \psi^{n+1,m+1} - \psi^{n+1,m}$, the finite difference alternating direction implicit formulation at every $\Delta t/3$ can be written as follow:

- Time splitting in z direction

$$\begin{aligned}
\left[\frac{S_s}{\phi}\theta_{i,j,k}^{n+\frac{1}{3},m}\right. &+ \left. C_{i,j,k}^{n+\frac{1}{3},m}\right]\frac{\delta_{i,j,k}^{m}}{(\Delta t/3)} - \frac{1}{\Delta z^2}\left[K_{i,j,k+\frac{1}{2}}^{n+\frac{1}{3},m}(\delta_{i,j,k+1}^{m} - \delta_{i,j,k}^{m}) - K_{i,j,k-\frac{1}{2}}^{n+\frac{1}{3},m}(\delta_{i,j,k}^{m} - \delta_{i,j,k-1}^{m})\right] \\
&= \frac{1}{\Delta x^2}\left[K_{i+\frac{1}{2},j,k}^{n+\frac{1}{3},m}(\psi_{i+1,j,k}^{n+\frac{1}{3},m} - \psi_{i,j,k}^{n+\frac{1}{3},m}) - K_{i-\frac{1}{2},j,k}^{n+\frac{1}{3},m}(\psi_{i,j,k}^{n+\frac{1}{3},m} - \psi_{i-1,j,k}^{n+\frac{1}{3},m})\right] \\
&+ \frac{1}{\Delta y^2}\left[K_{i,j+\frac{1}{2},k}^{n+\frac{1}{3},m}(\psi_{i,j+1,k}^{n+\frac{1}{3},m} - \psi_{i,j,k}^{n+\frac{1}{3},m}) - K_{i,j-\frac{1}{2},k}^{n+\frac{1}{3},m}(\psi_{i,j,k}^{n+\frac{1}{3},m} - \psi_{i,j-1,k}^{n+\frac{1}{3},m})\right] \\
&+ \frac{1}{\Delta z^2}\left[K_{i,j,k+\frac{1}{2}}^{n+\frac{1}{3},m}(\psi_{i,j,k+1}^{n+\frac{1}{3},m} - \psi_{i,j,k}^{n+\frac{1}{3},m}) - K_{i,j,k-\frac{1}{2}}^{n+\frac{1}{3},m}(\psi_{i,j,k}^{n+\frac{1}{3},m} - \psi_{i,j,k-1}^{n+\frac{1}{3},m})\right] \\
&+ \frac{K_{i,j,k+\frac{1}{2}}^{n+\frac{1}{3},m} - K_{i,j,k-\frac{1}{2}}^{n+\frac{1}{3},m}}{\Delta z} - \frac{S_s\theta_{i,j,k}^{n+\frac{1}{3},m}}{\phi(\Delta t/3)}\left[\psi_{i,j,k}^{n+\frac{1}{3},m} - \psi_{i,j,k}^{n}\right] - \frac{\theta_{i,j,k}^{n+\frac{1}{3},m} - \theta_{i,j,k}^{n}}{(\Delta t/3)}
\end{aligned} \tag{B5}$$

- Time splitting in x direction

$$
\begin{aligned}
\left[\frac{S_s}{\phi}\theta_{i,j,k}^{n+\frac{2}{3},m}\right. &+ \left. C_{i,j,k}^{n+\frac{2}{3},m}\right]\frac{\delta_{i,j,k}^{m}}{(\Delta t/3)} - \frac{1}{\Delta x^2}\left[K_{i+\frac{1}{2},j,k}^{n+\frac{2}{3},m}(\delta_{i+1,j,k}^{m} - \delta_{i,j,k}^{m}) - K_{i-\frac{1}{2},j,k}^{n+2/3,m}(\delta_{i,j,k}^{m} - \delta_{i-1,j,k}^{m})\right] \\
&= \frac{1}{\Delta x^2}\left[K_{i+\frac{1}{2},j,k}^{n+\frac{2}{3},m}(\psi_{i+1,j,k}^{n+\frac{2}{3},m} - \psi_{i,j,k}^{n+\frac{2}{3},m}) - K_{i-\frac{1}{2},j,k}^{n+\frac{2}{3},m}(\psi_{i,j,k}^{n+\frac{2}{3},m} - \psi_{i-1,j,k}^{n+\frac{2}{3},m})\right] \\
&+ \frac{1}{\Delta y^2}\left[K_{i,j+\frac{1}{2},k}^{n+\frac{2}{3},m}(\psi_{i,j+1,k}^{n+\frac{2}{3},m} - \psi_{i,j,k}^{n+\frac{2}{3},m}) - K_{i,j-\frac{1}{2},k}^{n+\frac{2}{3},m}(\psi_{i,j,k}^{n+\frac{2}{3},m} - \psi_{i,j-1,k}^{n+\frac{2}{3},m})\right] \\
&+ \frac{1}{\Delta z^2}\left[K_{i,j,k+\frac{1}{2}}^{n+\frac{2}{3},m}(\psi_{i,j,k+1}^{n+\frac{2}{3},m} - \psi_{i,j,k}^{n+\frac{2}{3},m}) - K_{i,j,k-\frac{1}{2}}^{n+\frac{2}{3},m}(\psi_{i,j,k}^{n+\frac{2}{3},m} - \psi_{i,j,k-1}^{n+\frac{2}{3},m})\right] \\
&+ \frac{K_{i,j,k+\frac{1}{2}}^{n+\frac{2}{3},m} - K_{i,j,k-\frac{1}{2}}^{n+\frac{2}{3},m}}{\Delta z} - \frac{S_s\theta_{i,j,k}^{n+\frac{2}{3},m}}{\phi(\Delta t/3)}\left[\psi_{i,j,k}^{n+\frac{2}{3},m} - \psi_{i,j,k}^{n+\frac{1}{3},m}\right] - \frac{\theta_{i,j,k}^{n+\frac{2}{3},m} - \theta_{i,j,k}^{n+\frac{1}{3}}}{(\Delta t/3)}
\end{aligned}
\tag{B6}
$$

- Time splitting in y direction

$$
\begin{aligned}
\left[\frac{S_s}{\phi}\theta_{i,j,k}^{n+1,m}\right. &+ \left. C_{i,j,k}^{n+1,m}\right]\frac{\delta_{i,j,k}^{m}}{(\Delta t/3)} - \frac{1}{\Delta y^2}\left[K_{i+\frac{1}{2},j,k}^{n+1,m}(\delta_{i+1,j,k}^{m} - \delta_{i,j,k}^{m}) - K_{i-\frac{1}{2},j,k}^{n+1,m}(\delta_{i,j,k}^{m} - \delta_{i-1,j,k}^{m})\right] \\
&= \frac{1}{\Delta x^2}\left[K_{i+\frac{1}{2},j,k}^{n+1,m}(\psi_{i+1,j,k}^{n+1,m} - \psi_{i,j,k}^{n+1,m}) - K_{i-\frac{1}{2},j,k}^{n+1,m}(\psi_{i,j,k}^{n+1,m} - \psi_{i-1,j,k}^{n+1,m})\right] \\
&+ \frac{1}{\Delta y^2}\left[K_{i,j+\frac{1}{2},k}^{n+1,m}(\psi_{i,j+1,k}^{n+1,m} - \psi_{i,j,k}^{n+1,m}) - K_{i,j-\frac{1}{2},k}^{n+1,m}(\psi_{i,j,k}^{n+1,m} - \psi_{i,j-1,k}^{n+1,m})\right] \\
&+ \frac{1}{\Delta z^2}\left[K_{i,j,k+\frac{1}{2}}^{n+1,m}(\psi_{i,j,k+1}^{n+1,m} - \psi_{i,j,k}^{n+1,m}) - K_{i,j,k-\frac{1}{2}}^{n+1,m}(\psi_{i,j,k}^{n+1,m} - \psi_{i,j,k-1}^{n+1,m})\right] \\
&+ \frac{K_{i,j,k+\frac{1}{2}}^{n+1,m} - K_{i,j,k-\frac{1}{2}}^{n+1,m}}{\Delta z} - \frac{S_s\theta_{i,j,k}^{n+1,m}}{\phi(\Delta t/3)}\left[\psi_{i,j,k}^{n+1,m} - \psi_{i,j,k}^{n+\frac{2}{3},m}\right] - \frac{\theta_{i,j,k}^{n+1,m} - \theta_{i,j,k}^{n+\frac{2}{3}}}{(\Delta t/3)}
\end{aligned}
\tag{B7}
$$

Similarly, the coupled Equations (B5), (B6), and (B7) are solved as three 1D problems for each directions of the 3D domain using the TDMA or PCR algorithms for tridiagonal matrices at one-third time step $\frac{\Delta t}{3}$. Right-hand sides of these equation consist of entirely known values at the time of computation.

Acknowledgement

We gratefully acknowledge support from Vietnam Education Foundation (VEF), Innovation grant from the Department of Civil and Environmental Engineering, Fellowship from the Computational Science and Engineering Program, and NSF Grants # EAR 1331906, CBET 1209402, EAR 1417444, and ICER 1440315

Code availability

Code available from server at: https://github.com/HydroComplexity/GCSFlow

References

References

Allen, M. B., & Murphy, C. (1985). A finite element collocation method for variably saturated flows in porous media. *Numerical Methods for Partial Differential Equations*, *1*, 229–239.

An, H., Ichikawa, Y., Tachikawa, Y., & Shiiba, M. (2011). A new iterative alternating direction implicit (iadi) algorithm for multi-dimensional saturated?unsaturated flow. *Journal of Hydrology*, *408*, 127 – 139.

Bennett, N. D., Croke, B. F., Guariso, G., Guillaume, J. H., Hamilton, S. H., Jakeman, A. J., Marsili-Libelli, S., Newham, L. T., Norton, J. P., Perrin, C., Pierce, S. A., Robson, B., Seppelt, R., Voinov, A. A., Fath, B. D., & Andreassian, V. (2013). Characterising performance of environmental models. *Environmental Modelling & Software*, *40*, 1 – 20.

Camporese, M., Daly, E., Dresel, P. E., & Webb, J. A. (2014). Simplified modeling of catchment-scale evapotranspiration via boundary condition switching. *Advances in Water Resources*, *69*, 95 – 105.

Camporese, M., Paniconi, C., Putti, M., & Orlandini, S. (2010). Surface-subsurface flow modeling with path-based runoff routing, boundary condition-based coupling, and assimilation of multisource observation data. *Water Resources Research*, *46*.

Celia, M. A., Bouloutas, E. T., & Zarba, R. L. (1990). A general mass-conservative numerical solution for the unsaturated flow equation. *Water Resources Research*, *26*, 1483–1496.

Dunne, T., Zhang, W., & Aubry, B. F. (1991). Effects of rainfall, vegetation, and microtopography on infiltration and runoff. *Water Resources Research*, *27*, 2271–2285.

Frei, S., & Fleckenstein, J. (2014). Representing effects of micro-topography on runoff generation and sub-surface flow patterns by using superficial rill/depression storage height variations. *Environmental Modelling & Software*, *52*, 5 – 18.

Frei, S., Lischeid, G., & Fleckenstein, J. (2010). Effects of micro-topography on surface-subsurface exchange and runoff generation in a virtual riparian wetland – a modeling study. *Advances in Water Resources*, *33*, 1388 – 1401.

Gasper, F., Goergen, K., Shrestha, P., Sulis, M., Rihani, J., Geimer, M., & Kollet, S. (2014). Implementation and scaling of the fully coupled terrestrial systems modeling platform (terrsysmp v1.0) in a massively parallel supercomputing environment ? a case study on juqueen (ibm blue gene/q). *Geoscientific Model Development*, *7*, 2531–2543.

van Genuchten, M. T. (1980). A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Science Society of America*, *44*, 892–898.

Gottardi, G., & Venutelli, M. (1993). A control-volume finite-element model for two-dimensional overland flow. *Advances in Water Resources*, *16*, 277 – 284.

Hammond, G. E., Lichtner, P. C., & Mills, R. T. (2014). Evaluating the performance of parallel subsurface simulators: An illustrative example with pflotran. *Water Resources Research*, *50*, 208–228.

Hockney, R. W., & Jesshope, C. R. (1988). *Parallel Computers Two: Architecture, Programming and Algorithms*. (2nd ed.). Bristol, UK: IOP Publishing Ltd.

Hromadka, T., & Lai, C. (1985). *Solving the two-dimensional diffusion flow model.* volume 1. New York, U.S.A., American Society of Civil Engineers.

Huyakorn, P. S., & Pinder, G. F. (1983). *Computational methods in subsurface flow*. Monograph Series. New York: Academic Press.

Hwang, H.-T., Park, Y.-J., Sudicky, E., & Forsyth, P. (2014). A parallel computational framework to solve flow and transport in integrated surface?subsurface hydrologic systems. *Environmental Modelling & Software*, *61*, 39 – 58.

Ivanov, V. Y., Bras, R. L., & Vivoni, E. R. (2008). Vegetation-hydrology dynamics in complex terrain of semiarid areas: 2. energy-water controls of vegetation spatiotemporal dynamics and topographic niches of favorability. *Water Resources Research*, *44*.

Ivanov, V. Y., Vivoni, E. R., Bras, R. L., & Entekhabi, D. (2004). Catchment hydrologic response with a fully distributed triangulated irregular network model. *Water Resources Research*, *40*.

Kirk, D. B., & Hwu, W.-m. W. (2010). *Programming Massively Parallel Processors: A Hands-on Approach*. (1st ed.). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Kollet, S. J., & Maxwell, R. M. (2006). Integrated surface–groundwater flow modeling: A free-surface overland flow boundary condition in a parallel groundwater flow model. *Advances in Water Resources*, *29*, 945 – 958.

Kollet, S. J., Maxwell, R. M., Woodward, C. S., Smith, S., Vanderborght, J., Vereecken, H., & Simmer, C. (2010). Proof of concept of regional scale hydrologic simulations at hydrologic resolution utilizing massively parallel computer resources. *Water Resources Research*, *46*.

Kumar, M., Duffy, C. J., & Salvage, K. M. (2009). A second-order accurate, finite volume–based, integrated hydrologic modeling (fihm) framework for simulation of surface and subsurface flow. *Vadose Zone Journal*, *8*, 873–890.

Kumar, P. (2011). Typology of hydrologic predictability. *Water Resources Research*, *47*.

Lal, A. (1998a). Performance comparison of overland flow algorithms. *Journal of Hydraulic Engineering*, *124*, 342–349.

Lal, A. (1998b). Weighted implicit finite-volume model for overland flow. *Journal of Hydraulic Engineering*, *124*, 941–950.

Le, P. V. V., & Kumar, P. (2014). Power law scaling of topographic depressions and their hydrologic connectivity. *Geophysical Research Letters*, *41*, 1553–1559.

Lee, V. W., Kim, C., Chhugani, J., Deisher, M., Kim, D., Nguyen, A. D., Satish, N., Smelyanskiy, M., Chennupaty, S., Hammarlund, P. et al. (2010). Debunking the 100x gpu vs. cpu myth: an evaluation of throughput computing on cpu and gpu. In *ACM SIGARCH Computer Architecture News* (pp. 451–460). ACM volume 38.

Lefsky, M. A., Cohen, W. B., Parker, G. G., & Harding, D. J. (2002). Lidar Remote Sensing for Ecosystem Studies. *BioScience*, (pp. 19–30).

Loos, M., & Elsenbeer, H. (2011). Topographic controls on overland flow generation in a forest -? an ensemble tree approach. *J. Hydrol.*, *409*, 94 – 103.

Maxwell, R. M. (2013). A terrain-following grid transform and preconditioner for parallel, large-scale, integrated hydrologic modeling. *Advances in Water Resources*, *53*, 109 – 117.

Maxwell, R. M., Chow, F. K., & Kollet, S. J. (2007). The groundwater–land-surface–atmosphere connection: Soil moisture effects on the atmospheric boundary layer in fully-coupled simulations. *Advances in Water Resources*, *30*, 2447–2466.

Maxwell, R. M., & Miller, N. L. (2005). Development of a coupled land surface and groundwater model. *Journal of Hydrometeorology*, *6*, 233–247.

Maxwell, R. M., Putti, M., Meyerhoff, S., Delfs, J.-O., Ferguson, I. M., Ivanov, V., Kim, J., Kolditz, O., Kollet, S. J., Kumar, M., Lopez, S., Niu, J., Paniconi, C., Park, Y.-J., Phanikumar, M. S., Shen, C., Sudicky, E. A., & Sulis, M. (2014). Surface-subsurface model intercomparison: A first set of benchmark results to diagnose integrated hydrology and feedbacks. *Water Resources Research*, *50*, 1531–1549.

Morita, M., & Yen, B. C. (2002). Modeling of conjunctive two-dimensional surface-three-dimensional subsurface flows. *Journal of Hydraulic Engineering*, *128*, 184–200.

Mualem, Y. (1976). A new model for predicting the hydraulic conductivity of unsaturated porous media. *Water Resources Research*, *12*, 513–522.

Nickolls, J., Buck, I., Garland, M., & Skadron, K. (2008). Scalable parallel programming with cuda. *Queue*, *6*, 40–53.

NVIDIA Corporation (2011). *NVIDIA CUDA C Programming Guide*.

Owens, J., Houston, M., Luebke, D., Green, S., Stone, J., & Phillips, J. (2008). Gpu computing. *Proceedings of the IEEE*, *96*, 879–899.

Panday, S., & Huyakorn, P. S. (2004). A fully coupled physically-based spatially-distributed model for evaluating surface/subsurface flow. *Advances in Water Resources*, *27*, 361 – 382. A Tribute to George F. Pinder.

Paniconi, C., & Putti, M. (1994). A comparison of picard and newton iteration in the numerical solution of multi-dimensional variably saturated flow problems. *Water Resources Research*, *30*, 3357–3374.

Paniconi, C., & Wood, E. F. (1993). A detailed model for simulation of catchment scale subsurface hydrologic processes. *Water Resources Research*, *29*, 1601–1620.

Qu, Y., & Duffy, C. J. (2007). A semidiscrete finite volume formulation for multiprocess watershed simulation. *Water Resources Research*, *43*.

Richards, L. A. (1931). Capillary conduction of liquids through porous mediums. *Physics*, *1*, 318–333.

Rodriguez-Iturbe, I. (2000). Ecohydrology: A hydrologic perspective of climate-soil-vegetation dynamics. *Water Resources Research*, *36*, 3–9.

Schwarz, B. (2010). Lidar: Mapping the world in 3d. *Nature Photonics*, *4*, 429–430.

Shen, C., & Phanikumar, M. S. (2010). A process-based, distributed hydrologic model based on a large-scale method for surface - subsurface coupling. *Advances in Water Resources*, *33*, 1524–1541.

Simmons, M. E., Ben Wu, X., & Whisenant, S. G. (2011). Plant and soil responses to created microtopography and soil treatments in bottomland hardwood forest restoration. *Restoration Ecology*, *19*, 136–146.

Sophocleous, M. (2002). Interactions between groundwater and surface water: the state of the science. *Hydrogeology Journal*, *10*, 52–67.

Sulis, M., Meyerhoff, S. B., Paniconi, C., Maxwell, R. M., Putti, M., & Kollet, S. J. (2010). A comparison of two physics-based numerical models for simulating surface water–groundwater interactions. *Advances in Water Resources*, *33*, 456–467.

Thomas, L. H. (1949). *Elliptic Problems in Linear Differential Equations over a Network*. Technical Report Columbia University.

Thompson, S. E., Katul, G. G., & Porporato, A. (2010). Role of microtopography in rainfall-runoff partitioning: An analysis using idealized geometry. *Water Resources Research*, *46*.

Ussyshkin, V., & Theriault, L. (2011). Airborne lidar: Advances in discrete return technology for 3d vegetation mapping. *Remote Sens.*, *3*, 416–434.

VanderKwaak, J., & Sudicky, E. (1999). Application of a physically-based numerical model of surface and subsurface water flow and solute transport. In F. Stauffer (Ed.), *ModelCARE 99 Proceedings of the International Conference on Calibration and Reliability in Groundwater Modeling: Coping with Uncertainty* (pp. 641–648). ETH Zurich, Zurich, Switzerland.

Vanka, S. P. (2013). 2012 Freeman Scholar Lecture: Computational Fluid Dynamics on Graphics Processing Units. *Journal of Fluids Engineering*, *135*, 061401+.

Weill, S., Mazzia, A., Putti, M., & Paniconi, C. (2011). Coupling water flow and solute transport into a physically-based surface–subsurface hydrological model. *Advances in Water Resources*, *34*, 128–136.

Winter, T. C., Harvey, J. W., Franke, O. L., & Alley, W. M. (1998). *Ground Water and Surface Water A Single Resource*. Technical Report 1139 US Geological Survey Circular.

Zhang, Y., Cohen, J., & Owens, J. D. (2010). Fast tridiagonal solvers on the gpu. *SIGPLAN Not.*, *45*, 127–136.

Table

Table 1: Parameters values for the Test Cases

Parameters	V-catchment	Infiltration Excess	Saturation Excess	Slab Case	Return Flow
Horizontal mesh size, $\Delta x = \Delta y$ $[m]$	5	20	20	1	5
Vertical mesh size, Δz $[m]$	n/a	0.2	0.2	0.05	0.05
Time step, Δt $[min]$	0.1	0.1	0.1	0.1	0.1
Initial water table depth, wt $[m]$	n/a	1.0	0.5, 1.0	1.0	0.5
Specific storage, S_s $[m^{-1}]$	n/a	5×10^{-4}	5×10^{-4}	5×10^{-4}	5×10^{-4}
Porosity, ϕ [-]	n/a	0.4	0.4	0.4	0.4
Saturated hydraulic conductivity, K_{sat} $[m\ min^{-1}]$	n/a	6.94×10^{-5} 6.94×10^{-6}	6.94×10^{-4}	6.94×10^{-4} 6.94×10^{-6}	6.94×10^{-2}
Manning's coefficients, n_b $[m^{-\frac{1}{3}}\ min]$	n/a				
- Hillslope	2.5×10^{-4}	3.31×10^{-4}	3.31×10^{-4}	3.31×10^{-4}	3.31×10^{-4}
- Channel	2.5×10^{-3}	na	na	na	na
Rainfall rate $[m\ min^{-1}]$	1.8×10^{-4}	3.3×10^{-4}	3.3×10^{-4}	3.3×10^{-4}	1.5×10^{-4}
Evaporation rate $[m\ min^{-1}]$	0	0	0	0	5.4×10^{-6}
x-direction slope, S_x [%]	5.0	0.05	0.05	0.05	0.5, 5
y-direction slope, S_y [%]	2.0	0	0	0	0
vanGenuchten parameters					
Alpha, α $[cm^{-1}]$	n/a	1.0	1.0	1.0	1.0
Pore-size distribution, n [-]	n/a	2.0	2.0	2.0	2.0
Residual water content, θ_r [-]	n/a	0.08	0.08	0.08	0.08
Saturated water content, θ_s [-]	n/a	0.4	0.4	0.4	0.4

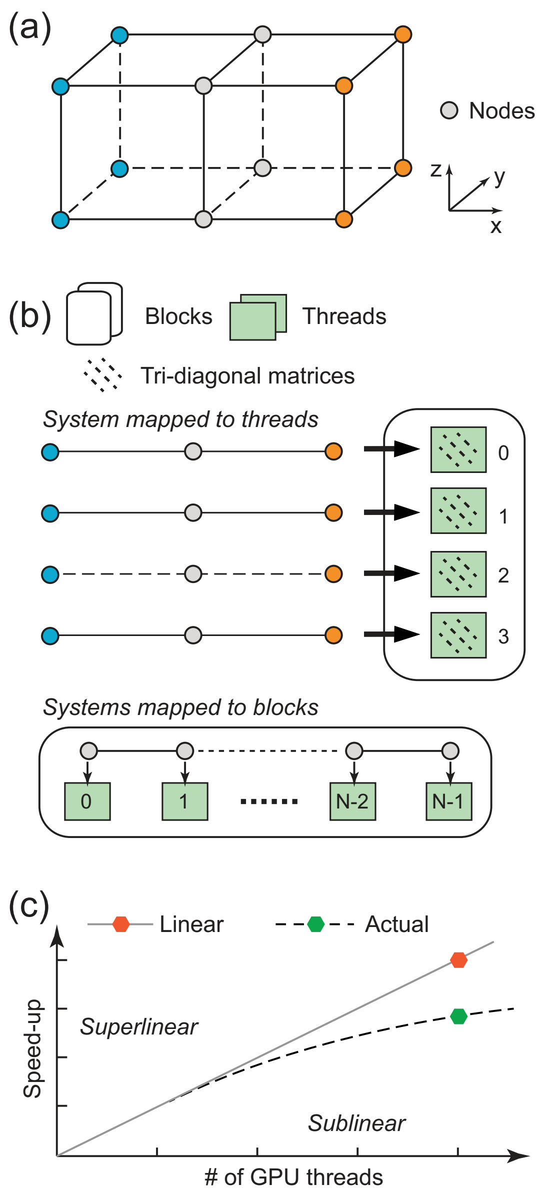

Figure 1: Parallel implementation of the ADI solvers in GPU computing. (a) Computational grid domain in three-dimensional space. (b) ADI approach discretizes the computational domain into 1-D problems involving independent tridiagonal linear systems. Each is assigned into a single thread (system mapped to threads) or block (system mapped to block) for numerical solution. A large number of fine-grained GPU processors can solve these systems in parallel. (c) Illustration of actual speed-up in GPU parallel computing. Thread synchronization is required at all time step, which reduces the speed-up from a linear trend for iterative ADI solver.

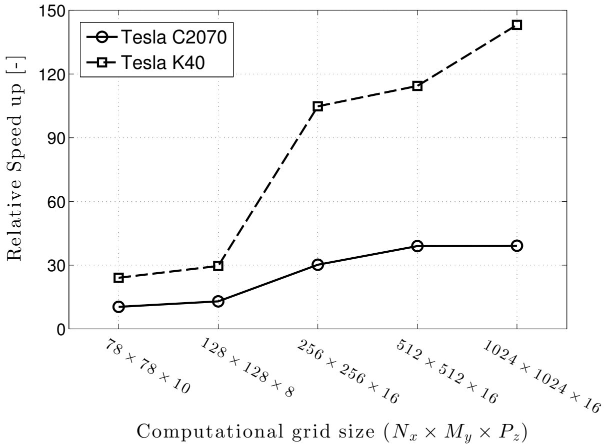

Figure 2: Relative speedup of the iterative ADI solver for tridiagonal matrix systems at different sizes of problem between GPUs (C2070 and K40) and CPUs (Xeon 2.67GHz). The speedup is the ratio between simulation time in CPU and in each GPU.

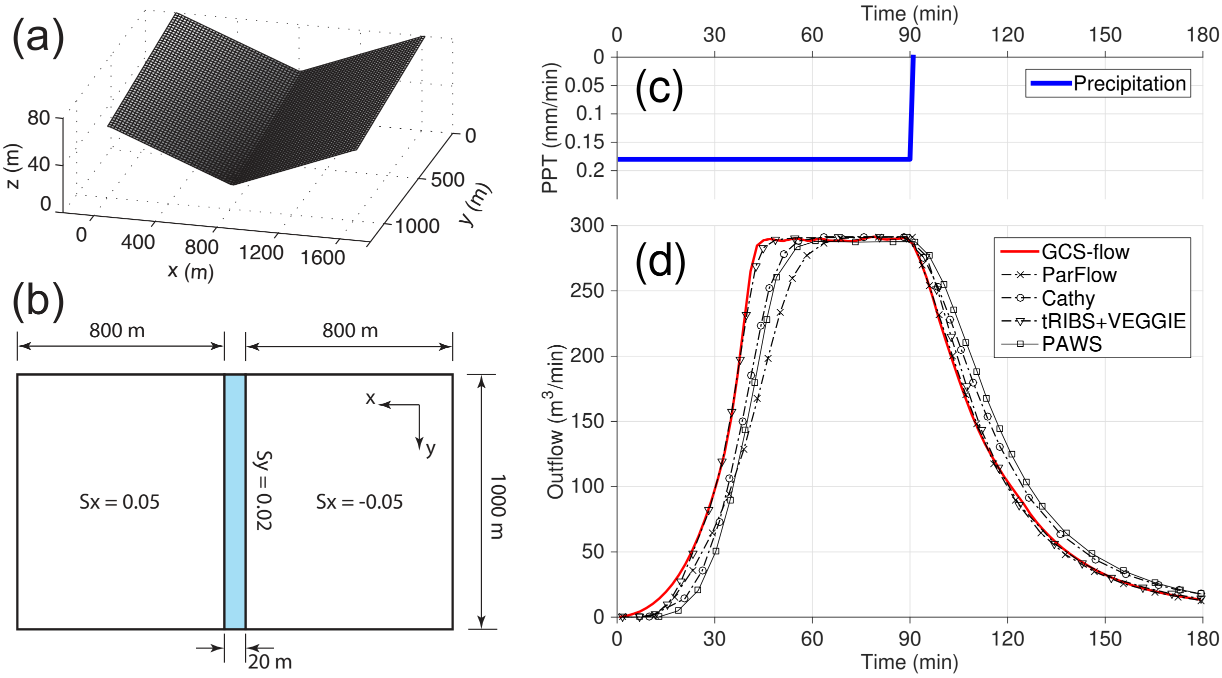

Figure 3: Benchmarking outflow response using a tilted V-catchment for overland and channel flow [after Sulis et al. (2010); Maxwell et al. (2014)]. (a) Tilted V-catchment domain - three-dimensional view. (b) Tilted V-catchment domain - top view. (c) Rainfall series consists of a uniform rainfall event from 0 to 90 [min] followed by 90 [min] of no rainfall. (d) Comparisons of overland and channel outflow at the outlet between GCS-flow and other models.

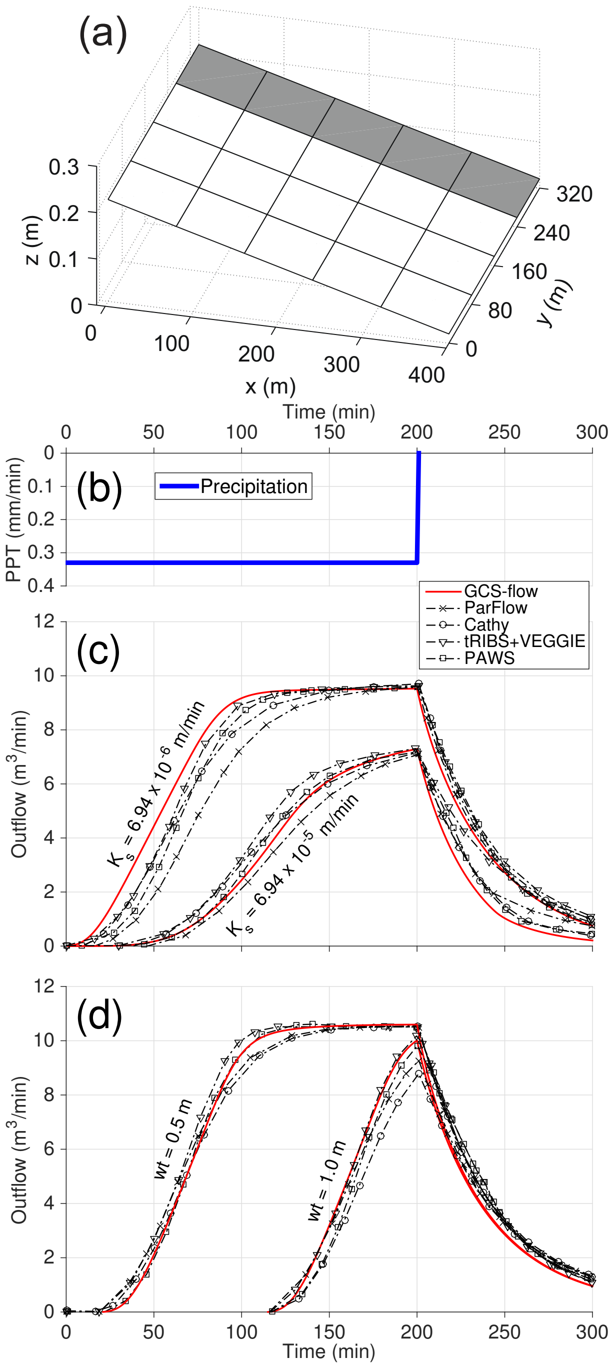

Figure 4: Benchmarking outflow response using homogeneous sloping tests [after Sulis et al. (2010); Maxwell et al. (2014)] (a) Illustration of the domain for sloping tests. Outflow is measured at the outlet of the grey strip of cells. (b) Rainfall series consists of a uniform rainfall event from 0 to 200[min] followed by 100 [min] of no rainfall. (c and d) Comparisons of outflow at the outlet between GCS-flow and other models for two test cases: (c) Infiltration excess with two different values of hydraulic conductivity; and (d) Saturation excess with two different values of water table depth.

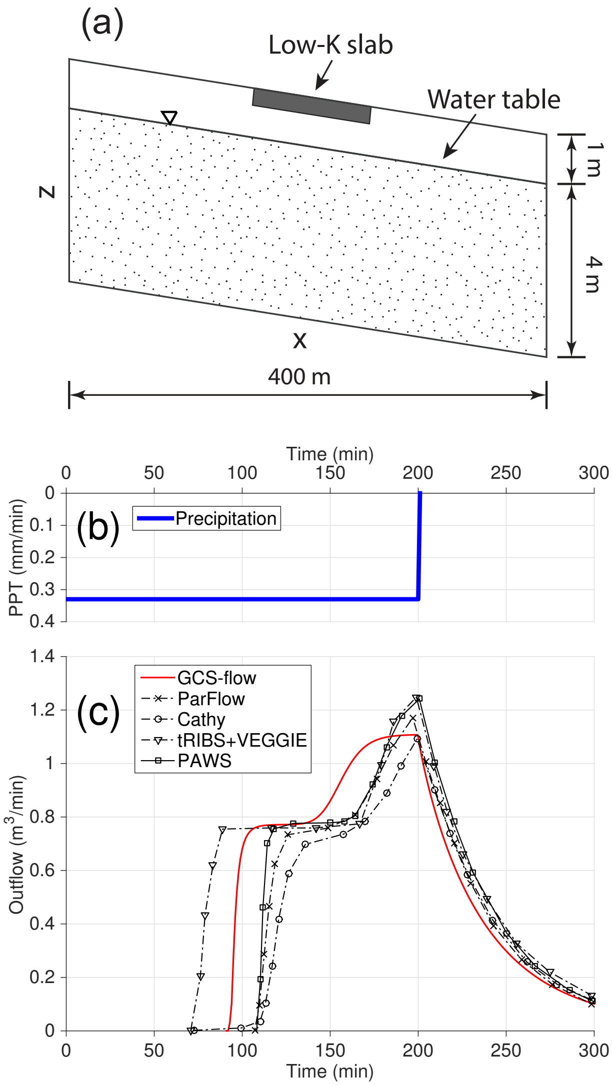

Figure 5: Benchmarking flow response using heterogeneous (slab) sloping tests [after Sulis et al. (2010); Maxwell et al. (2014)] (a) Domain and hydraulic conductivity distribution for slab test. (b) Rainfall series consists of a uniform rainfall event from 0 to 200 [min] followed by 100 [min] of no rainfall. (c)Comparison of outflow at the outlet between GCS-flow and other models.

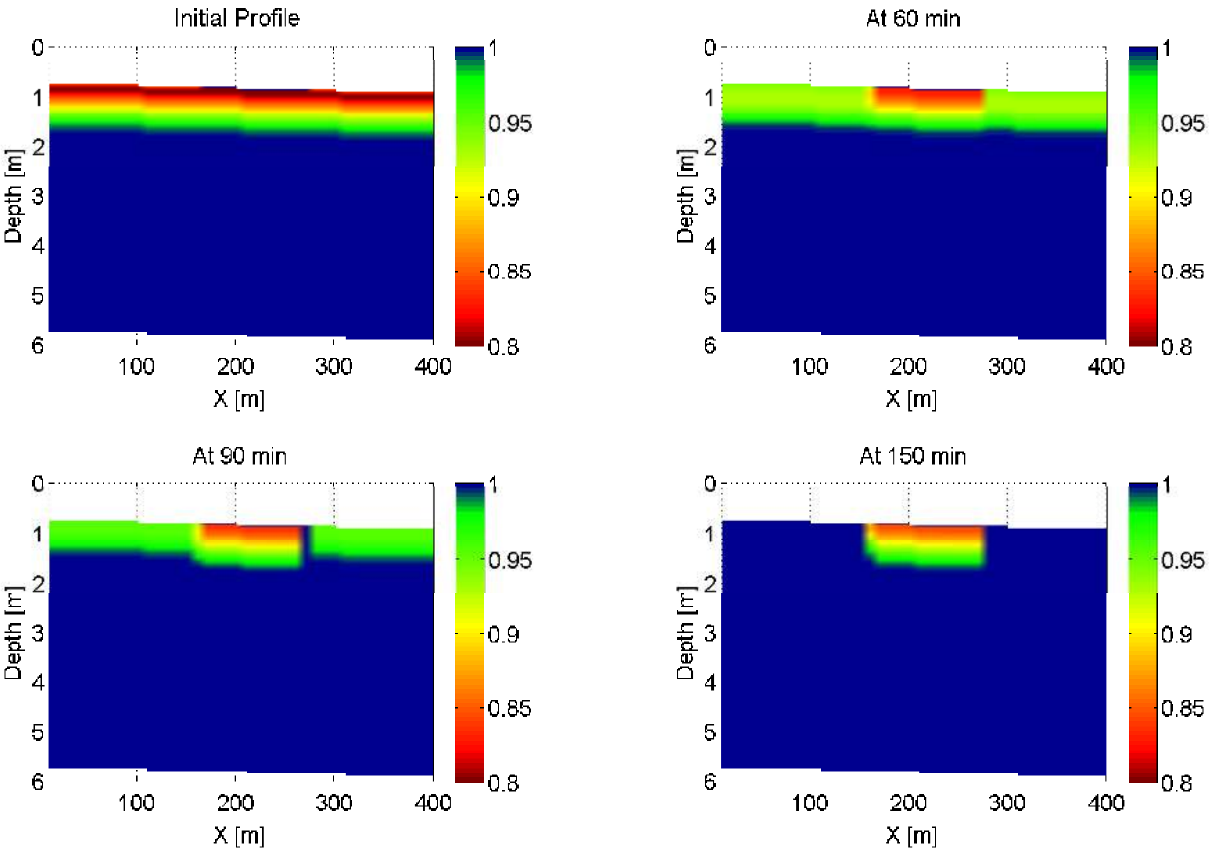

Figure 6: Saturation profile for the slab test at time 0, 60, 90, and 150 [min] obtained from GCS-flow model

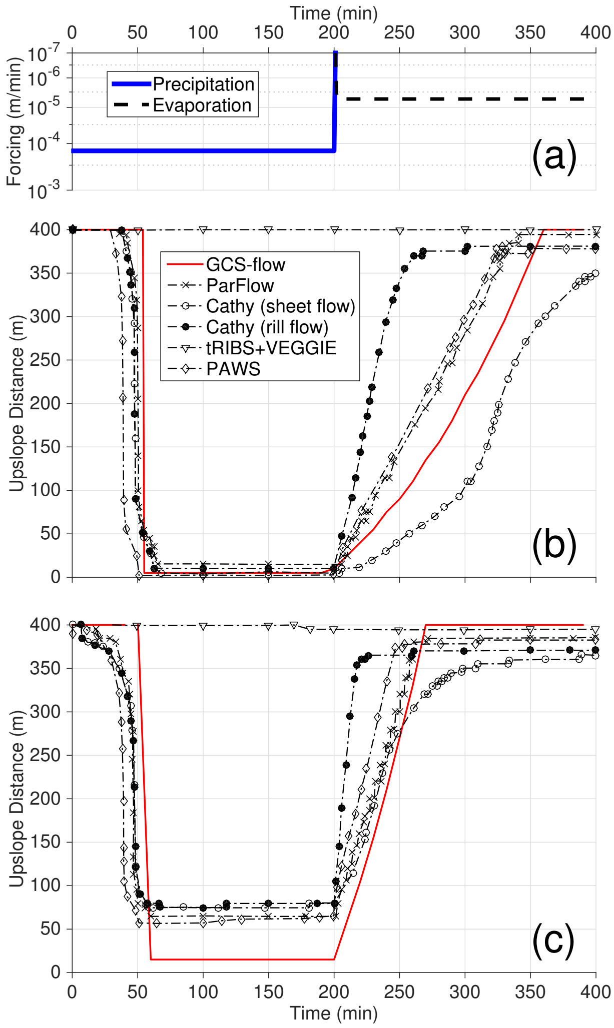

Figure 7: Benchmarking the evolution of the intersection point between the water table and the land surface [after Sulis et al. (2010); Maxwell et al. (2014)]. (a) Atmospheric forcing consist of a uniform rainfall event from 0 to 200 [min] followed by 200 [min] of uniform evaporation in log scale. (b, c) Simulation results for the return flow test using hillslope inclination of (b) 0.5% slope and (c) 5 % slope.

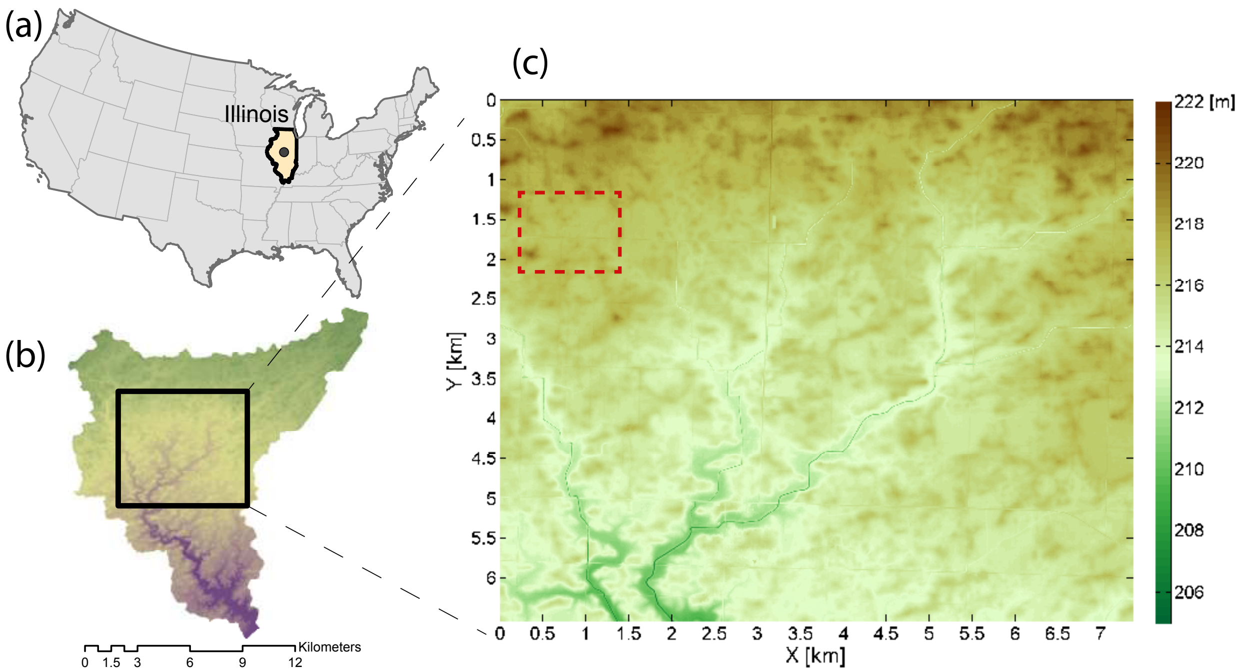

Figure 8: Area of simulation used to illustrate the GCS-flow model at lidar-data resolution. (a and b) Map of Goose Creek watershed of the Sangamon River Basin in central Illinois, USA. (c) Lidar data in Goose Creek watershed. Red rectangle shows the area used in illustration in Figure 9.

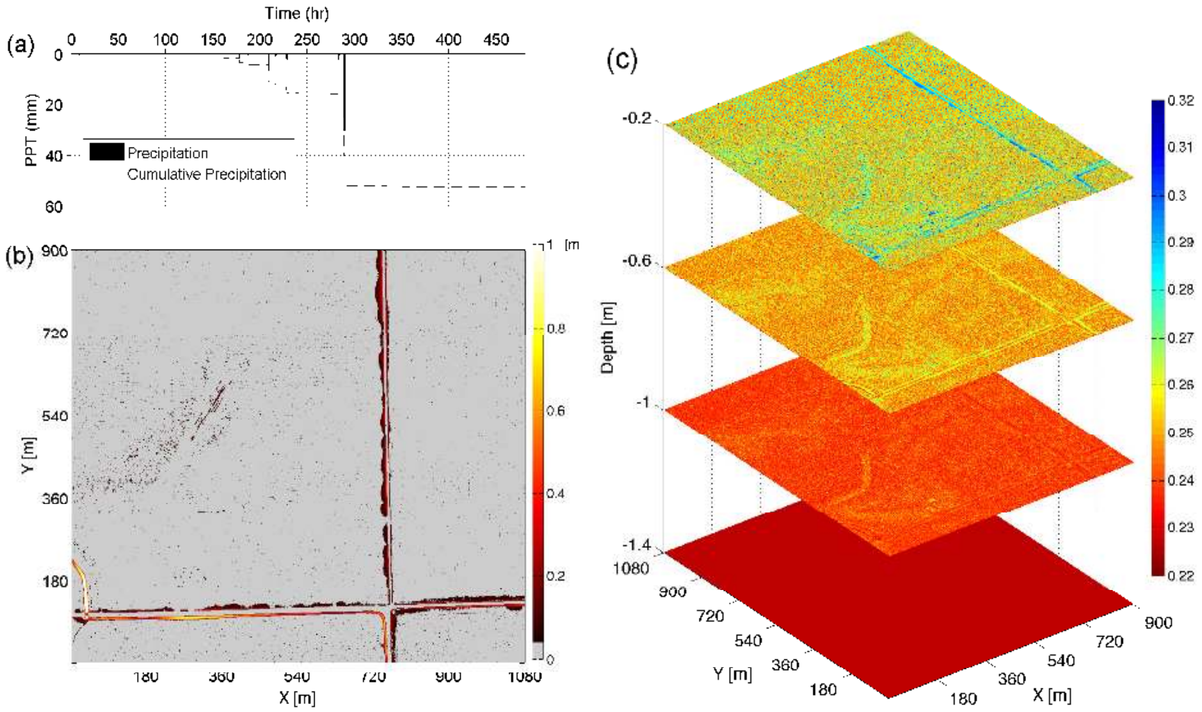

Figure 9: Profile of the spatial distribution of soil-moisture for the rectangle domain shown in Figure 8c. (a) Precipitation time series used for the simulation period; (b) Water depth simulated on the study area by GCS-flow at 320 hr; and (c) Corresponding soil moisture profile at layers over depth simulated by GCS-flow at 320 hr