AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE

10.1002/2016WR019191

The integrated hydrologic model intercomparison project, IH-MIP2: A second set of benchmark results to diagnose integrated hydrology and feedbacks

Stefan Kollet[1,2], **Mauro Sulis**[3], **Reed Maxwell**[4], **Claudio Paniconi**[5], **Mario Putti**[6], AQ7
Giacomo Bertoldi[7], **Ethan T. Coon**[8], **Emanuele Cordano**[7,9], **Stefano Endrizzi**[10], **Evgeny Kikinzon**[8],
Emmanuel Mouche[11], **Claude Mügler**[11], **Young-Jin Park**[12], **Jens Christian Refsgaard**[13],
Simon Stisen[13], and **Edward Sudicky**[14]

[1]Agrosphere Institute, Forschungszentrum Jülich GmbH, Germany, [2]Centre for High-Performance Scientific Computing in Terrestrial Systems, HPSC TerrSys, Geoverbund ABC/J, Jülich, Jermany, [3]Meteorological Institute, Bonn University, Germany, [4]Department of Geology and Geological Engineering, Colorado School of Mines, Mines, Colorado, USA, [5]Institut National de la Recherche Scientifique, Centre Eau Terre Environnement, Université du Québec, Canada, [6]Department of Mathematics, University of Padova, Italy, [7]Institute for Alpine Environment, EURAC, European Academy Bolzano, Italy, [8]Computational Earth Science, Los Alamos National Laboratory, Los Alamos, New Mexico, USA, [9]Rendena100, Engineering and Consultancy sole proprietorship, Tione di Trento, Italy, [10]Department of Geography, University of Zurich, Switzerland, [11]Laboratoire des Sciences du Climat et de l'Environnement, CEA-CNRS-UVSQ, France, [12]Aquanty, Inc., Waterloo, Canada, [13]Department of Hydrology, Geological Survey of Denmark and Greenland, Copenhagen, Denmark, [14]Department of Earth and Environment, University of Waterloo, Waterloo, Canada AQ1

Key Points:
- Seven hydrologic models were intercompared using three benchmarks of increasing complexity
- Models showed good agreement with respect to various hydrologic responses (storage, discharge, and soil moisture values)
- Different discretizations of the digital elevation model had stronger influence than mathematical model formulation

Correspondence to:
S. Kollet,
s.kollet@fz-juelich.de

Citation:
Kollet, S, M Sulis, R Maxwell, C Paniconi, M Putti, G Bertoldi, ET Coon, E Cordano, S Endrizzi, E Kikinzon, E Mouche, C Mügler, Y-J Park, J Christian Refsgaard, S Stisen, and E Sudicky. (2016), The integrated hydrologic model intercomparison project, IH-MIP2: A second set of benchmark results to diagnose integrated hydrology and feedbacks, *Water Resour. Res.*, *52*, doi:10.1002/2016WR019191.

Received 10 MAY 2016
Accepted 28 NOV 2016
Accepted article online 17 DEC 2016

Abstract Emphasizing the physical intricacies of integrated hydrology and feedbacks in simulating connected, variably saturated groundwater-surface water systems, the Integrated Hydrologic Model Intercomparison Project initiated a second phase (IH-MIP2), increasing the complexity of the benchmarks of the first phase. The models that took part in the intercomparison were ATS, Cast3M, CATHY, GEOtop, HydroGeoSphere, MIKE-SHE, and ParFlow. IH-MIP2 benchmarks included a tilted v-catchment with 3-D subsurface; a superslab case expanding the slab case of the first phase with an additional horizontal subsurface heterogeneity; and the Borden field rainfall-runoff experiment. The analyses encompassed time series of saturated, unsaturated, and ponded storages, as well as discharge. Vertical cross sections and profiles were also inspected in the superslab and Borden benchmarks. An analysis of agreement was performed including systematic and unsystematic deviations between the different models. Results show generally good agreement between the different models, which lends confidence in the fundamental physical and numerical implementation of the governing equations in the different models. Differences can be attributed to the varying level of detail in the mathematical and numerical representation or in the parameterization of physical processes, in particular with regard to ponded storage and friction slope in the calculation of overland flow. These differences may become important for specific applications such as detailed inundation modeling or when strong inhomogeneities are present in the simulation domain.

1. Background and Introduction

With the advent of a number of integrated hydrologic modeling systems [*Ebel et al.*, 2009], *Maxwell et al.* [2014] identified the need for a formalized Integrated Hydrologic Model Intercomparison Project (IH-MIP), in order to inform the hydrologic science community on the status of hydrologic model development. There is continued scientific interest in understanding complex interactions between variably saturated groundwater and surface water flow especially under heterogeneous conditions, when nonlinear hydrodynamics across various space and time scales influence hydrologic response. The mathematical representation of these interactions in simulation models is still a great challenge, because of the composite physical processes described by nonlinear, coupled partial differential equations that cannot be validated easily in the classical sense (i.e., comparison with analytical solutions), in the case of realistic flow problems. Thus, uncertainty remains in the attribution of hydrologic responses (e.g., correspondence to actual processes) to model

AGU Water Resources Research 10.1002/2016WR019191

structural errors (e.g., missing processes, such as water use), and initial and boundary conditions (e.g., complex domains) in the evaluation with in situ and remotely sensed observations.

The basic ideas of model intercomparison have been pursued in many studies across a number of Earth science disciplines leading to a more complete understanding of model physics and parameterizations and increased confidence in simulation results [e.g., *Bowling et al*., 2003; *Smith and Gupta*, 2012; *Steefel et al*., 2015; *Taylor et al*., 2012]. The participants of the IH-MIP identified a comparative approach based on the juxtaposition of results from simulations performed with different hydrologic models for a number of numerical experiments with increasing complexity and realism. The approach is based on the rationale that a comparative approach with increasing complexity utilizing, first, simplified synthetic numerical experiments, and, second, real-world catchments in conjunction with observations is useful in order to establish a baseline of understanding of the impact of numerical couplings (e.g., groundwater-surface water, groundwater-vadose zone) and the representation of heterogeneity in hydraulic properties. This baseline is required before including complex land surface processes, such as evaporation from bare soil and root water uptake by plants, in the intercomparison.

Following the positive experience and outcome of the first Phase of the IH-MIP reported in *Maxwell et al*. [2014], a second Phase (IH-MIP2) was launched with a workshop at the Center for High Performance Scientific Computing in Terrestrial Systems, Geoverbund ABC/J in Bonn, Germany in June 2013. The goal of IH-MIP2 was to progress from purely 2-D horizontal overland flow and 2-D vertical groundwater-surface water coupling with simple heterogeneity to (i) fully 3-D groundwater-surface water coupling, (ii) more complex heterogeneity, and (iii) a field experiment published previously *Abdul and Gillham* [1989], thereby moving toward more realistic catchment processes, geometries, and scales. In the following sections, the participating integrated hydrologic model codes are briefly introduced, including models that have joined the IH-MIP since Phase 1. Detailed descriptions of three test cases are provided so as to allow for the reproduction of the simulations and results with most available integrated hydrologic modeling tools. The results are provided and discussed in the context of process represention and model couplings including an analysis of agreement.

2. Model Descriptions

2.1. Participating Hydrologic Models

Seven models took part in IH-MIP2: ATS, Cast3M, CATHY, GEOtop, HydroGeoSphere (HGS), MIKE-SHE, and ParFlow (PF). These are introduced briefly below including key references for the interested reader.

2.1.1. ATS

ATS (Advanced Terrestrial Simulator) is a collection of ecosystem hydrology process models built on top of the Amanzi modeling platform and the Arcos multiphysics management strategy [*Coon et al*., 2016]. ATS can solve problems for thermal hydrology in both the surface and subsurface, including freeze/thaw processes, a surface energy balance, and snow processes including depth hoar and aging. ATS also uses a simple big-leaf model to incorporate dynamic vegetation and carbon cycling, includes some simple deformation capabilities, and can solve problems with reactive transport through Amanzi. Here ATS is used to solve Richards' equation in the subsurface and a diffusive wave model for the surface; these are coupled through a continuous pressure formulation. ATS uses mimetic finite differences on unstructured meshes to maintain accuracy through high aspect ratio cells and layering structures typical of hydrogeology applications [*Brezzi et al*., 2005]. ATS was not part of the first intercomparison, but has solved those problems as part of model validation.

2.1.2. Cast3M

Cast3M is a simulation platform developed at the French Alternative Energies and Atomic Energy Commission (CEA) in France. It is devoted to solid and fluid mechanics problems in research and engineering. The platform offers computational, preprocessing (mesh generation), and postprocessing (visualization) functionalities. Cast3M can solve hydrology and hydrogeology problems (flow and transport) either in finite elements or finite volumes. The coupling of surface and subsurface flows is performed within a Darcy multidomain approach [*Weill et al*., 2009]. Surface runoff is solved in a 3-D porous layer, called runoff layer, which is added at the top of the subsurface model. For the three test cases, the cells are quadrilateral in both the surface and subsurface domains and follow the terrain at the topographic slope of the surface. The equations are discretized with a finite volume scheme employing upwind and cell-centered fluxes at

Water Resources Research 10.1002/2016WR019191

the surface and in the subsurface, respectively. This approach unifies the Darcy and Richards equations (subsurface) with the diffusive wave approximation of the Saint Venant equations for surface flows (runoff and streams) into a single generalized Richards equation with domain-dependent parameters and physical laws. This equation is solved with an implicit time scheme. The nonlinear terms are calculated with an iterative Picard algorithm and an underrelaxation method for the nonlinear parameters that depend on water pressure. A multidomain transport equation (advection, diffusion, dispersion) is also coupled with the generalized Richards equation for simulating tracer problems. It allows tracking of event and preevent water during a rainfall event, for instance. The Darcy multidomain approach developed in Cast3M has been applied to 2-D and 3-D configurations [*Weill et al*., 2009] and to test cases of the first Phase of the IH-MIP, although Cast3M was not part of the first exercise.

2.1.3. CATHY

CATHY (CATchment HYdrology) [*Bixio et al*., 2002; *Camporese et al*., 2010] solves the integrated model by coupling a finite element approach for the three-dimensional Richards equation with a finite difference discretization of a path-based 1-D kinematic approximation of the Saint Venant equation. Surface-subsurface coupling is addressed on the basis of a time-splitting procedure that iteratively updates boundary conditions to automatically partition potential fluxes (rainfall and evapotranspiration) into actual fluxes across the land surface. Mass balance equations are then used to evaluate changes in surface and subsurface storage. This procedure ensures that pressure and flux continuity is enforced at the surface/subsurface interface. Important innovations to the model with respect to Phase 1 include coupling CATHY with the Noah-MP land surface model [*Niu et al*., 2014a,b], incorporating detailed vegetation models coupled with simplified boundary layer dynamics [*Bonetti et al*., 2015; *Manoli et al*., 2014, 2015], and adding coupled hydrogeophysical inversion via data assimilation [*Manoli et al*., 2014; *Rossi et al*., 2015].

2.1.4. GEOtop

GEOtop [*Endrizzi et al*., 2014; *Rigon et al*., 2006] is a grid-based distributed hydrological model that describes three-dimensional water flow in the soil and at the soil surface, as well as water and energy exchanges with the atmosphere, considering vegetation processes and the effects of complex topography on radiative fluxes. A snow multilayer model and soil freezing and thawing processes are integrated [*Dall'Amico et al*., 2011]. Vegetation dynamics is optionally simulated with an external module [*Della Chiesa et al*., 2014]. The heat and water flow equations are linked in a time-lagged manner [e.g., *Panday and Huyakorn*, 2004], with a three-dimensional finite volume approach solved by a special Newton-Raphson method, where the grid is slope-normal in order to allow a proper description of mass and energy exchange processes in steep terrain. Unsaturated and saturated zones are solved in the same equation system: when the soil is unsaturated, the water content is calculated with the soil water retention curve according to the *van Genuchten* [1980] formula, whereas in case of saturated zones, the linear concept of specific storativity is used. The surface (or overland) water flow is described with the approximation proposed by *Gottardi and Venutelli* [1993]. GEOtop was not part of first intercomparison project.

2.1.5. HGS

HGS (HydroGeoSphere) is a 3-D control-volume, finite element simulator designed to simulate the entire terrestrial portion of the hydrologic cycle [*Aquanty*, 2015]. It uses a globally implicit approach to simutaneously solve the diffusive wave equation for surface water flow and Richards' equation for subsurface flow. It dynamically integrates the key components of the hydrologic cycle, such as evaporation from bare soil and surface water bodies, vegetation-dependent transpiration with the dynamics of changes in leaf area, root density and root depth, snow accumulation and melt, and soil freeze and thaw. Features such as macropores, fractures, extraction wells, and tile drains can either be incorporated discretely or using a dual-porosity dual-permeability formulation. As with the solution of the coupled water flow equations, HGS solves the contaminant and energy transport equations over the land surface and in the subsurface, thus allowing for surface/subsurface interactions. Atmospheric interactions for an energy balance can be parameterized and solved within the HGS platform [*Brookfield et al*., 2009] or HGS can be coupled with the Weather Research and Forcast (WRF) model for a seamless simulation of atmosphere, surface, and subsurface interactions [*Davison et al*., 2015]. The HGS platform uses a Newton method combined with an iterative sparse matrix solver to handle nonlinearities in the governing flow equations. It has been parallelized to utilize high-performance computing facilities to address large-scale problems [*Hwang et al*., 2014].

2.1.6. MIKE-SHE

MIKE-SHE is a flexible software package for modeling the major processes in the hydrologic cycle and includes models for evapotranspiration, overland flow, unsaturated flow, groundwater flow, and channel

AGU Water Resources Research 10.1002/2016WR019191

flow [*Abbott et al*., 1986; *Butts et al*., 2004]. The modeling system has been used worldwide for both commercial and scientific applications across a range of scales and water-related issues [*Larsen et al*., 2016; *Wijesekara et al*., 2014]. The flexibility of the system allows the user to combine process descriptions and numerical solutions to fit the problem at hand [*Graham and Butts*, 2005]. Of specific interest to the current study are the saturated and unsaturated process descriptions and their coupling. For the saturated zone, variations of the hydraulic head are described mathematically by the 3-D Darcy equation and discretized numerically by an iterative implicit finite difference technique solved by the preconditioned conjugate gradient (PCG) method [*Graham and Butts*, 2005]. Unsaturated flow is simulated using a fully implicit finite difference solution to the Richards equation [*Refsgaard and Storm*, 1995]. Unsaturated flow is considered only as 1-D in the vertical direction and therefore ignores any horizontal flow. The saturated and unsaturated zones are linked by an explicit coupling and solved in parallel, instead of being solved in a single matrix [*Storm*, 1991]. The advantage of explicit coupling is that the time stepping for the unsaturated and saturated zones can be different, reflecting their difference in time scales [*Graham and Butts*, 2005]. This makes the code computationally attractive compared to the more complete single matrix solutions at the cost of a simplification to 1-D unsaturated flow and the risk of instability of the coupling scheme. It should be noted that the MIKE-SHE modeling system is designed for catchment scale models ($10^{\circ}-10^5$ km^2), where lateral fluxes are small compared to vertical fluxes in the unsaturated zone. In MIKE-SHE, overland flow is included via the diffusive wave approximation using a 2-D finite difference approach. The presented model simulations with MIKE-SHE were carried out by the Geological Survey of Denmark and Greenland, who are users of the MIKE-SHE modelling software without access to the source code and who are not the model developers.

2.1.7. PF

PF (ParFlow) is a 3-D variably saturated groundwater-surface water flow model that treats the groundwater, vadose zone, and surface water as a single continuum based on the Richards and Saint Venant equations. The system of coupled equations is solved in a finite control volume approach with two-poi[*Kollet and Maxwell*, 2008]nt flux approximation in a globally implicit implementation using a regular grid [*Jones and Woodward*, 2001; *Kollet and Maxwell*, 2006]. In this study, the saturation and relative permeability relationships are parameterized using the van Genuchten equation [*van Genuchten*, 1980]. PF has been integrated with land surface processes and subsurafce energy transport [*Kollet and Maxwell*, 2008; *Kollet et al*., 2009; *Maxwell and Miller*, 2005], and various atmospheric models [*Maxwell et al*., 2007, 2011; *Shrestha et al*., 2014] in order to close the terrestrial hydrdologic and energy cycle from groundwater across the landsurface into the atmosphere. In addition, the terrain following grid (not applied in IH-MIP2) in PF [*Maxwell*, 2013] affords large-scale high-resolution simulations at the continental scale [*Maxwell et al*., 2015]. In PF, the solution algorithms and preconditioners were shown to exhibit excellent parallel efficiency [*Gasper et al*., 2014; *Kollet et al*., 2010; *Osei-Kuffuor et al*., 2014]. Recently, PF was incorporated with the Parallel Data Assimilation Framework [*Kurtz et al*., 2016] affording efficient state and parameter estimation.

2.2. Key Distinctions of the Numerical Representations of Physical Processes

Some major distinctions in the representation of physical processes in the different models are summarized in the following paragraphs. These are important in the interpretation and discussion of the results in the ensuing sections.

2.2.1. Treatment of the Saturated-Unsaturated Zone

Most of the applied models (ATS, Cast3M, CATHY, GEOtop, HGS, PF) are continuum models treating the saturated and unsaturated zones as well as the surface water flow domain as a single continuum in three spatial dimensions (Table 1). In case of saturation, the concept of specific storage is applied. In MIKE-SHE, the coupling between the unsaturated and saturated zones is solved by an iterative mass balance procedure, in which the lower nodes of the unsaturated compartment are solved separately in a pseudo time step. The mass-conservative solution is achieved by using a stepwise adjustment of the water table and recalculation of the solution for the unsaturated zone. The iterative procedure conserves the mass for the entire column by considering outflows and source/sink terms in the saturated zone.

2.2.2. Coupling of Variably Saturated Groundwater-Surface Water Flow

ATS, Cast3M, and PF apply a free surface overland flow boundary condition at the top (i.e., the land surface) based on pressure and flux continuity at the surface (Table 1). Thus, no interface between the surface and subsurface flow domains described by a conductance concept is assumed. For the coupling of surface and subsurface flow equations in HGS, the continuity of pressure can be enforced across the surface and

AQ6

Table 1. Summary of Key Distinctions and Similarities of Physical Representations in the Seven Models of This Study

Model	Saturated-Unsaturated	Discretization Scheme	Coupling Subsurface-Surface	Surface Storage/Flow	Heterogeneity Representation
ATS	Continuum	Mimetic finite differences	Free surface BC	Diffusive wave	Fully distributed
Cast3M	Continuum	Finite volume, quadrilateral	Free surface BC	Diffusive wave	Fully distributed
CATHY	Continuum	Finite element	BC switching	Diffusive wave	Fully distributed
GEOtop	Continuum	Finite differences, rectangular	Free surface BC	Kinematic wave	Soil classes, profiles, bedrock properties
HGS	Continuum	Finite element, quadrilateral	First-order exchange	Diffusive wave	Fully distributed
MIKE-SHE	Coupled	Finite difference, rectangular	Information not available	Diffusive wave	Fully distributed
PF	Continuum	Finite control volume, rectangular	Free surface BC	Kinematic wave	Fully distributed

subsurface domains or a first-order exchange formulation can be utilized for flux continuity [*Liggett et al.*, 2012]. In this study, a first-order exchange formulation was applied. In CATHY, flux and pressure continuity at the surface/subsurface interface is enforced by means of a boundary condition switching procedure commonly used in variably saturated subsurface flow models to track atmosphere-controlled (Neumann boundary condition at the land surface) and soil-limited (Dirichlet condition) infiltration and evaporation dynamics. This procedure is extended to the integrated model by allowing for ponding at the surface. In case of MIKE-SHE, information on the coupling was not provided by the developers at the date of publication.

2.2.3. Surface Storage and Surface Water Flow

ATS solves a diffusive wave approximation and also uses Manning's roughness approach for calculating the flow law; no surface storage or rill parameterization are included in these runs (Table 1). CATHY allows for depression storage and uses rill-based routing that is parameterized dynamically and independently for overland and channel flow paths [*Orlandini and Rosso*, 1998]. GEOtop use also a parameterization based on a Manning's-type equation which allows surface ponding in depressions and below a user-defined rill storage height similar to HGS. In HGS, surface water flow is simulated based on the diffusive wave approximation and a modified Manning's equation: it is assumed that surface water can flow laterally only once water levels are above a rill storage height (depression storage) and it slowly approaches to the full flow capacity after water levels exceed the obstruction storage (e.g., vegetation) height. Note, however, that rill and obstruction storages were not applied for the HGS benchmark simulations in this study. In MIKE-SHE, the diffusive wave approximation is also applied using a 2-D finite-difference approach including a Strickler/Manning-type approach with an optional surface detention storage. In PF, no surface storage or rills are parameterized, and surface water flow is simulated based on the kinematic wave approximation including Manning's roughness approach and friction slopes for each grid cell. The same approach is adopted in Cast3M except for the use of the diffusive wave approximation.

2.2.4. Subsurface Heterogeneity

In ATS, Cast3M, CATHY, HGS, MIKE-SHE, and PF subsurface heterogeneity can be implemented in a fully distributed way with cell or element-wise, spatially varying hydraulic properties (Table 1). In GEOtop, subsurface heterogeneity can be defined by a variable number of soil classes and profiles and in terms of bedrock depth and properties.

3. Benchmark Simulations, Phase 2

The second set of benchmark simulations were first published online at www.hpsc-terrsys.de and underwent successive revisions to facilitate a constructive intercomparison. The benchmarks consist of a tilted v-catchment, already used in Phase 1 for overland flow only, this time with coupled 3-D groundwater-surface water flow in recession and rainfall-recession experiments; a superslab experiment derived from the slab experiment of Phase 1, this time with additional layered heterogeneity intersecting the land surface at a short distance from the hillslope outlet; and a simulation of the Borden field experiment consisting of a rainfall-runoff experiment along a ditch on the order of 80 m length. Note that each modeling group performed numerical convergence tests for the respective benchmarks in order to provide the best available solution. These solutions were obtained by making sure that the sequence formed by the Euclidian norms of the differences between two successively refined runs was indeed converging to zero. Because of the different types of models and the different types of computational grids, the metric by which this difference is

Water Resources Research 10.1002/2016WR019191

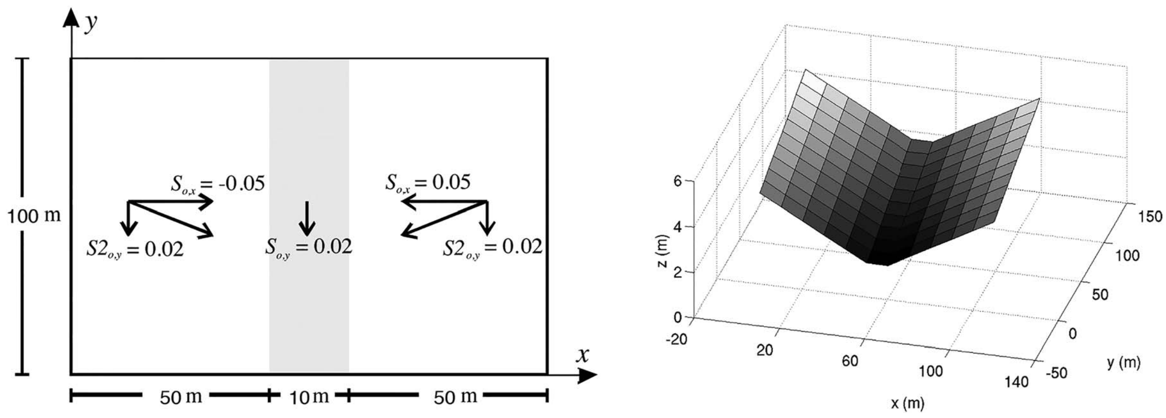

Figure 1. (left) 2-D and (right) 3-D schematic of the tilted v-catchment benchmark.

evaluated differs from code to code. Nonetheless, it gives the necessary confidence that the code is converging toward an asymptotic state as the mesh parameters decrease.

3.1. The 3-D Tilted v-Catchment

The 3-D tilted v-catchment benchmark (Figure 1) expands upon the prior surface-flow only case and now extends into the subsurface. This benchmark consists of two identical hillslopes with uniform topographic slope and a channel in the center of the domain. The subsurface extends 5 m below the land surface and is homogeneous in all hydrogeologic properties (Table 2). The simulation models for this benchmark were initialized with vertically hydrostatic initial conditions and a water table 2 m below the land surface. The total simulation period was 120 h. Two different scenarios were simulated encompassing a recession and a rainfall-recession scenario (Table 3). The hydraulic conductivity of the subsurface was assigned a large value to obtain a quick response in case of scenario 1. In order to obtain a ratio between the precipitation rate and hydraulic conductivity that is not too small [*Maxwell and Kollet*, 2008], the precipitation rate was fixed at 0.1m/h in the case of scenario 2. The spatial discretization and time stepping varied between the different models (Table 4).

Table 2. Model Geometry, Initial and Boundary Conditions, and Hydraulic Parameters for the Tilted v-Catchment Benchmark

Model geometry	
Lateral extensions in x and y:	110×100 m
Vertical extension in z:	5 m below land surface
Lateral and vertical resolutions:	Varies between models
Boundary conditions	
Overland flow:	Critical depth
Subsurface lateral and bottom:	No flow
Subsurface top:	Overland flow
Initial conditions	
Water table (hydraulic pressure, $p = 0$m) 2 m below land surface, hydrostatic conditions vertically	
Hydraulic parameters overland flow:	
Friction slope in x direction:	$S_{f,x} = 0.05$ hillslopes; $S_{f,x} = 0.0$ channel
Friction slope in y direction:	$S_{f,y} = 0.02$ everywhere
Manning's roughness hillslope:	$n_{hs} = 1.74 \times 10^{-4}$ h/m$^{1/3}$
Manning's roughness channel:	$n_c = 1.74 \times 10^{-3}$ h/m$^{1/3}$
Hydraulic parameters subsurface	
Saturated hydraulic conductivity:	$K_{sat} = 10.0$ m/h
van Genuchten rel. permeability:	$n = 2.0$ and $\alpha = 6.0$ m^{-1}
Res. and sat. vol. water content:	$\theta_{res} = 0.08$ and $\theta_{sat} = 0.4$
Porosity:	$\phi = 0.4$
Specific storage:	$S_s = 1.0 \times 10^{-5}$ m^{-1}
Simulation period and time stepping	
Simulation period:	120 h
Time step size:	Variable between models

3.2. Superslab

The overall geometry of the superslab benchmark described in *Kollet and Maxwell* [2006] and *Maxwell et al.* [2014] is made more complex here with an additional layered, low-conductivity heterogeneity relatively close to the hillslope outlet intersecting the land surface (Figure 2). The subsurface extends to 5 m below the land surface. The simulation was initialized with vertically hydrostatic conditions and a water table 5 m below the land surface everywhere. A single scenario was simulated consisting of 12 h total simulation time with 3 h of rain followed by 9 h of recession. Parameter values, boundary and initial conditions, and timing information are summarized in Table 5. Again, the spatial discretization

AGU Water Resources Research 10.1002/2016WR019191

Table 3. Scenario Information for the Tilted v-Catchment Benchmark

Scenario 1 (S1):	
No rainfall; return flow only based on initial conditions	
Scenario 2 (S2):	
Rainfall	
Rain duration:	20 h
Rain rate:	$q_r = 0.1$ m/h
Recession duration:	100 h

and time stepping varied between the different models (Table 6).

3.3. Borden Benchmark

The Borden test case is based on the original field experiment and hydraulic parameters of *Abdul and Gillham* [1989] and consists of a ditch of approximately 2 m depth that was uniformly irrigated with water containing a dilute bromide solution for 50 min (Figure 3). The spatial extent of the domain was approximately 20 m $\times$ 80 m in the horizontal direction with an arbitrary horizontal base (or datum) at 0 m. Two digital elevation models (DEMs) at 0.5 m and 1 m resolution were provided and are available online at www.hpsc-terrsys.de in simple ASCII format. Here results for the 0.5 m resolution are shown. Required information on boundary conditions and hydraulic properties are listed in Table 7. The simulation period included the aforementioned 50 min of rainfall followed by 50 min of recession, i.e., 100 min total simulation time. As with the other two benchmarks, the spatial discretization and time stepping for the Borden case varied between the different models (Table 8).

4. Analysis of Agreement Between Models

In benchmarking numerical models, the true solution is often not known and, thus, there is no simulation result that can be used as a reference, in order to decide, if a model is better than another. Therefore, only the relative agreement between models can be assessed. In the study by *Duveiller et al*. [2016], commonly used metrics of agreement have been discussed. We choose in particular the Pearson product-moment correlation coefficient

$$r=\frac{n^{-1}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sigma_X\sigma_Y} \tag{1}$$

where $\bar{X}$ and $\bar{Y}$ are the mean values of the data sets (vectors) X and Y, respectively, and σ_X and σ_Y are their standard deviations.

In equation (1), r is a measure of the linear agreement/dependence of X and Y ranging between -1 and 1, and is equivalent to the coefficient of determination in the case of a linear regression model. In this study, because time series of simulations that describe the same dynamic process are being compared, r reflects how well two different models agree, for that given process or response variable, in terms of their temporal deviations with respect to their mean responses. However, r does not provide any insight into the agreement of absolute values and, thus, into potential additive and multiplicative biases when models diverge. Based on an index by *Mielke* [1991], *Duveiller et al*. [2016] proposed a new metric

$$\lambda=\alpha r \tag{2}$$

where the coefficient

$$\alpha=\frac{2}{\sigma_X/\sigma_Y+\sigma_Y/\sigma_X+(\bar{X}-\bar{Y})^2/(\sigma_X\sigma_Y)} \quad \text{for } r>0,\ \text{otherwise}\ \alpha=0 \tag{3}$$

represents any bias (additive/multiplicative) between the two data sets and ranges between 1 (no bias, perfect agreement) and 0 (full bias, no agreement). Thus,

Table 4. Summary of the Spatial and Temporal Discretization of the Different Models for the Tilted v-Catchment Benchmark

Model	Horizontal Resolution (m)	Vertical Resolution (m)	Temporal Resolution (s)min, mean, max
ATS	2.5	0.125	Adaptive S1: 1, 165, 1800 S2: 1, 34, 265
Cast3M	S1: 2.5	0.1	Adaptive 2, 10, 20
	S2: 5.0	$0.0005<\Delta z<0.4$	Adaptive 0.01, 2, 30
CATHY	10	0.05	Adaptive S1: 60, 60, 60 S2: 3.7, 5.0, 8.0
GEOtop	10	0.05	Constant 10
HGS	S1: 1.0	0.25	Adaptive 0.01, 581, 600
	S2: 5.0	0.1	Adaptive 0.01, 581, 600
MIKESHE	1	0.1	Adaptive 0.21, 2.3, 180
PF	1	0.05	Constant 6

AGU Water Resources Research 10.1002/2016WR019191

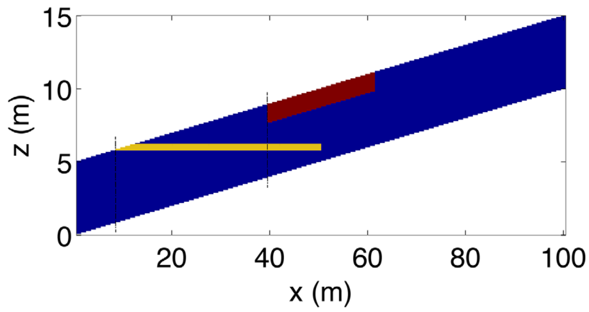

Figure 2. Schematic of the superslab benchmark with two heterogeneous slabs (slab 1 in yellow and slab 2 in red) in a homogeneous matrix (blue). The vertical-dashed lines show the locations of the cross sections plotted in Figure 11.

Duveiller et al. [2016] scale the correlation coefficient r with a factor that accounts for systematic differences between the two data sets. Note that $\lambda=0$ when $r \leq 0$, i.e., negatively correlated data sets are considered to have no agreement (equations (14) and (15) in *Duveiller et al.* [2016]).

In our analysis, instead of presenting λ values directly, α and r were calculated in a pairwise fashion for all combinations of models and expressed graphically in a matrix. This provides a differentiated picture of the agreement between models in terms of the temporal dynamics with respect to the mean behavior (r values) and of the presence of any biases between the different models (α values).

5. Results and Discussion

In the analyses of the results, emphasis was placed on different storage terms (saturated, unsaturated, ponded) and discharge at different locations of the domain. Profiles and cross sections at characteristic times and locations were compared in order to identify and understand local differences for the

Table 5. Model Geometry, Initial, and Boundary Conditions, Hydraulic Parameters, and Simulation Periods for the 2-D Vertical Superslab Benchmark

Model geometry				
Lateral extensions in x:				100 m
Vertical extension in z:				5 m below land surface
First slab, lateral extension in x:				8–50 m
First slab, lateral extension in z:				5.8–6.2 m
Second slab, lateral extension in x:				40–60 m
Second slab, lateral extension in z:				1.3 m below the land surface
Boundary conditions				
Overland flow:				Critical depth
Subsurface lateral & bottom:				No flow
Subsurface top:				Overland flow
Initial conditions				
Water table (hydraulic pressure, $p=0$ m) 5 m below land surface, hydrostatic conditions vertically				
Hydraulic parameters—overland flow:				
Friction slope in x direction:				$S_{f,x}=0.1$
Manning's roughness:				$n_c=1.0\times10^{-6}$ h/m$^{1/3}$
	K_{sat} (m/h)	Porosity, ϕ		Specific storage, S_s (m^{-1})
Hydraulic parameters—subsurface				
Domain	10.0	0.1		1.0×10^{-5}
First slab	0.025	0.1		1.0×10^{-5}
Second slab	0.001	0.1		1.0×10^{-5}
		α (m^{-1})	θ_{res}	θ_{sat}
Van Genuchten parameters				
Domain	2.0	6.0	0.02	0.1
First slab	3.0	1.0	0.03	0.1
Second slab	3.0	1.0	0.03	0.1
Simulation period				
Simulation period:			12 h	
Rain duration:			3 h	
Rain rate:			$q_r=0.05$ m/h	
Recession duration			9 h	

Table 6. Summary of Discretization Schemes and Spatial and Temporal Discretization of the Different Models for the Superslab Benchmark

Model	Horizontal Resolution (m)	Vertical Resolution (m)	Temporal Resolution (s)min, mean, max
ATS	1	0.05	Adaptive 1.6, 16, 60
Cast3M	2	$3.0 \times 10^{-5} < \Delta z < 0.05$	Adaptive 10^{-4}, 2, 30
CATHY	1	0.05	Adaptive 0.1, 11, 60
GEOtop	1	0.05	Constant 9
HGS	1	0.05	Adaptive 3.6×10^{-3}, 144, 180
MIKESHE	1	0.05	Adaptive 1.1, 1.7, 3.6
PF	1	0.05	Constant 6

heterogeneous superslab and Borden benchmarks. In order to obtain a quantitative picture of the agreement between different models, the analysis outlined in section 4 was performed in a pairwise fashion and the results are presented as matrices.

5.1. Tilted v-Catchment

Figures 4 and 5 show the storage and discharge time series of recession scenario S1 and the results of the analysis of agreement. In this scenario, the catchment approaches hydrostatic conditions starting from the initial condition due to gravity drainage. Thus, the water table, which initially follows the land surface, equilibrates horizontally leading to an intersection with the land surface and discharge at the catchment outlet. Because the dynamics are quite subtle and slow, especially with respect to ponding of water at the land surface, this is a challenging problem to simulate.

For the case of unsaturated and saturated storage, there is a relatively strong intermodel variability until 20 h of simulation time, however the absolute differences are small (about 7% in the case of unsaturated storage), which is reflected in relatively small *r* values, in the case of unsaturated storage (Figure 5). After 20 h simulation, there is a clear difference in the trend of the recession for ATS, Cast3M, CATHY, HGS, and PF compared to MIKE-SHE and GEOtop. In the former five models, which are all based on the continuum approach, unsaturated storage increases, while in MIKE-SHE, unsaturated storage decreases, resulting in small α values in Figure 5. The pronounced increase in unsaturated storage in the case of GEOtop leads to negative correlations with the other models and, thus, $\alpha = 0$.

The continued decrease of unsaturated storage during the recession phase could be explained by the 1-D simplification of the vadose zone in MIKE-SHE or by the unsaturated-saturated zone coupling. The 1-D simplification ignores any horizontal redistribution between unsaturated columns, thus reducing unsaturated storage over time via leakage from the unsaturated compartment into the groundwater compartment. As for GEOtop, the jagged recession behavior of the unsaturated storage is likely due to the fact that relatively thick soil layers (50 mm) switch from saturated to almost saturated conditions several times during the simulation. In general, saturated storages agree reasonably well between the different models, with MIKE-SHE and GEOtop providing the smallest α values (Figure 5).

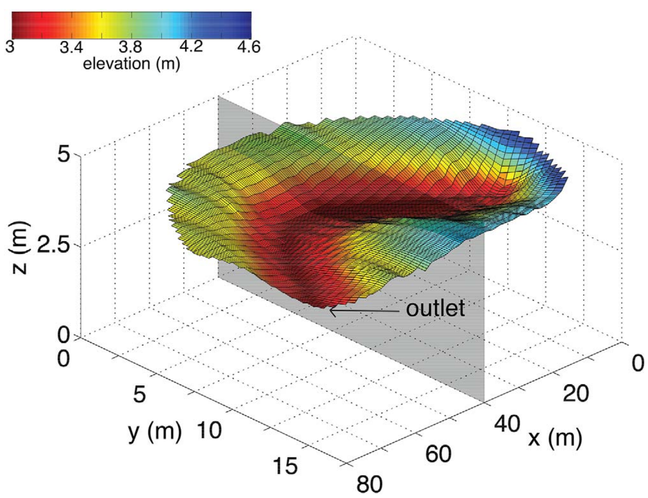

Figure 3. Topography of the Borden domain and location of the cross section shown in Figure 14.

While the temporal dynamics agree quite well (large *r* values), absolute ponded storages differ by more than a factor of two between the different models, reflected in the small α values in Figure 5. In Cast3M, ponded storage is very sensitive to surface mesh refinement. MIKE-SHE shows a noisier output, which is also very sensitive to the mesh. In the aforementioned numerical convergence study, the ponded storage becomes asymptotically smaller with finer vertical discretization at the

Table 7. Model Geometry, Initial, and Boundary Conditions, Hydraulic Parameters, and Simulation Periods for the Borden Benchmark

Model geometry	
Approximately 20 m x 80 m; ditch with 2 m depth; datum at 0 m; max. elevation approximately 4.64 m	
DEM, 0.5 m resolution:	dem0.5m.dat
Boundary conditions	
Overland flow:	Critical depth everywhere
Subsurface lateral & bottom:	No flow
Subsurface top:	Overland flow
Initial conditions	
Water table 20 cm below ditch outlet ($z = 2.78$ m above datum), vertically hydrostatic conditions	
Hydraulic parameters overland flow	
Manning's roughness:	$n = 8.3 \times 10^{-5}$ h/m$^{1/3}$
Hydraulic parameters subsurface flow	
Saturated hydraulic conductivity:	$K = 0.036$ m/h
van Genuchten:	$n = 6$ and $\alpha = 1.9$ m^{-1}
Res. and sat. vol. water content:	$\theta_{res} = 0.067$ and $\theta_{sat} = 0.37$
Porosity:	$\phi = 0.37$
Specific storage:	$S_s = 3.2 \times 10^{-4}$ m^{-1}
Simulation period, time stepping and scenarios	
Simulation period:	100 min
Rain duration:	50 min
Rain rate:	$q_r = 0.02$ m/h
Recession duration:	50 min

surface, similar to results reported in *Kollet and Maxwell* [2006]. However, absolute values of ponded storage are small compared to total storage values. Nonetheless, the differences may be significant in case of inundation modeling, where minor changes in topography may lead to large differences in ponded area and storage. Discharges again agree quite well between the different continuum models, reflected in large r and α values (Figure 5). MIKE-SHE simulates higher discharge values, which may be attributed to the increased drainage from the vadose zone and increased saturated storage. GEOtop also simulates higher discharge, which is coherent with the estimation of high ponded storage, implying a high water level at the outlet.

In Figure 5, the continuum models show a pattern of agreement (large r and α values) for the subsurface storages and discharge, with the exception of GEOtop for unsaturated storage. Ponded storage shows generally the smallest α values, suggesting that this storage term is generally well captured in terms of temporal dynamics and less in terms of absolute values.

In the second scenario, S2, the pure recession response of S1 is superposed by a rainfall event during the first 20 h of the simulation. With respect to the unsaturated and saturated storage time series (Figure 6), this leads to a distinct separation of the continuum models (ATS, Cast3M, CATHY, GEOtop, HGS, PF) and the compartment model (MIKE-SHE). The continuum models predict generally lower unsaturated and larger-saturated storages compared to MIKE-SHE. The observed differences are significant in case of unsaturated storage (up to a factor of six), resulting in small α values in Figure 7 and less significant in the case of saturated storage (less than a factor of two). The differences can again be explained by the 1-D simplification of the unsaturated zone in MIKE-SHE. In the continuum models, a horizontal flux can be generated between partially saturated cells which enables a faster downhill water movement and thereby higher saturated storage at the expense of

Table 8. Summary of Discretization Schemes and Spatial and Temporal Discretization of the Different Models for the Borden Benchmark

Model	Horizontal Resolution (m)	Vertical Resolution (m)	Temporal Resolution (s)min, mean, max
ATS	0.5	$0.05 \leq \Delta z \leq 0.628$	Adaptive 0.16, 5.2, 14.6
Cast3M	0.5	$0.001 \leq \Delta z \leq 1$	Adaptive 0.001, 1, 300
CATHY	0.5	$0.015 \leq \Delta z \leq 0.15$	Adaptive 3.6×10^{-3}, 0.4, 3.0
GEOtop	0.5	$0.01 \leq \Delta z \leq 0.1$	Constant 180
HGS	0.5	$0.15 \leq \Delta z \leq 0.45$	Adaptive 0.5, 50, 60
MIKESHE	0.5	0.01	Adaptive 1.7×10^{-3}, 7.5×10^{-3}, 60,
PF	0.5	0.05	Constant 5

AGU Water Resources Research 10.1002/2016WR019191

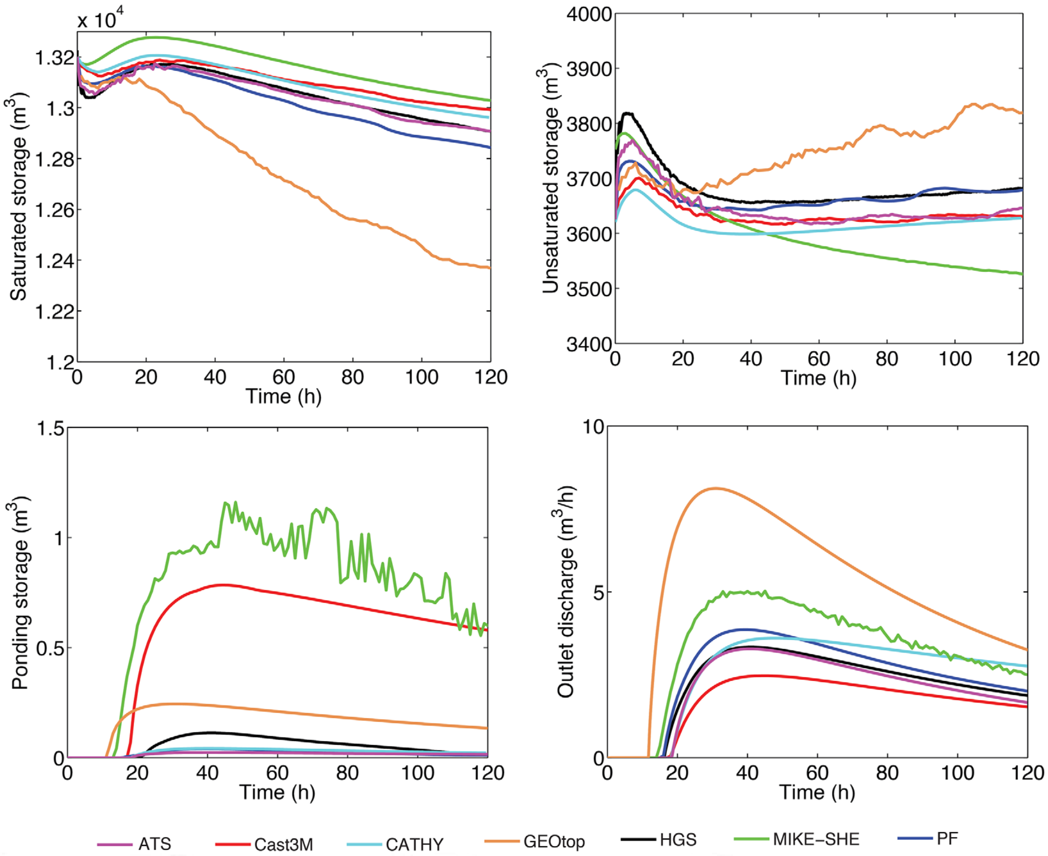

Figure 4. Storage and discharge time series of scenario S1 of the tilted v-catchment benchmark: saturated, unsaturated, ponded storage, and discharge at the outlet.

unsaturated storage. In MIKE-SHE, horizontal water transfer remains inactive until a given column is saturated.

The constant rainfall rate applied was smaller than the saturated hydraulic conductivity, thus, excess saturation is the sole process of runoff generation in the simulation. It is notable that all models provide almost identical discharge time series (all r and α values are close to 1), while the ponded storage at the land surface may differ by some 30%, yet biases are small (Figures 6 and 7). In Figure 7, all models are part of a pattern of good agreement (high r and α values) for all variables, except MIKE-SHE in the case of unsaturated storage, similar to the performance obtained for test case SC1. A decrease in the values for CATHY is also detectable, due to too much ponding and saturated storage.

All models arrive at steady state after some 10 h of simulation time and also exhibit remarkable agreement during the recession period, which is due to the strong excitation of the models by the relatively strong rainfall event of 100 mm/h. This lends confidence in the models' ability to consistently simulate rainfall-runoff responses during and after strong rainfall and the process of saturation excess when most of the catchment area contributes to runoff. The models, however, implement different overland flow and surface storage parameterizations, leading to the differences in ponded storage at the surface, which may again be important in inundation modeling. These parameterizations are relatively straightforward to implement and modify, and may be unified and tested for consistency between different modeling platforms if required.

5.2. Superslab

In the superslab benchmark (Figures 8 and 9), a gravitational equilibration of the laterally nonhydrostatic initial condition is superposed with a 3 h rainfall event producing a complex series of interactions of variably saturated groundwater flow and surface runoff. Here, excess infiltration and saturation interact at the slabs, producing local ponding, runon and runoff, and regions of excess saturation (Figure 10). Given the complexity of the interactions,

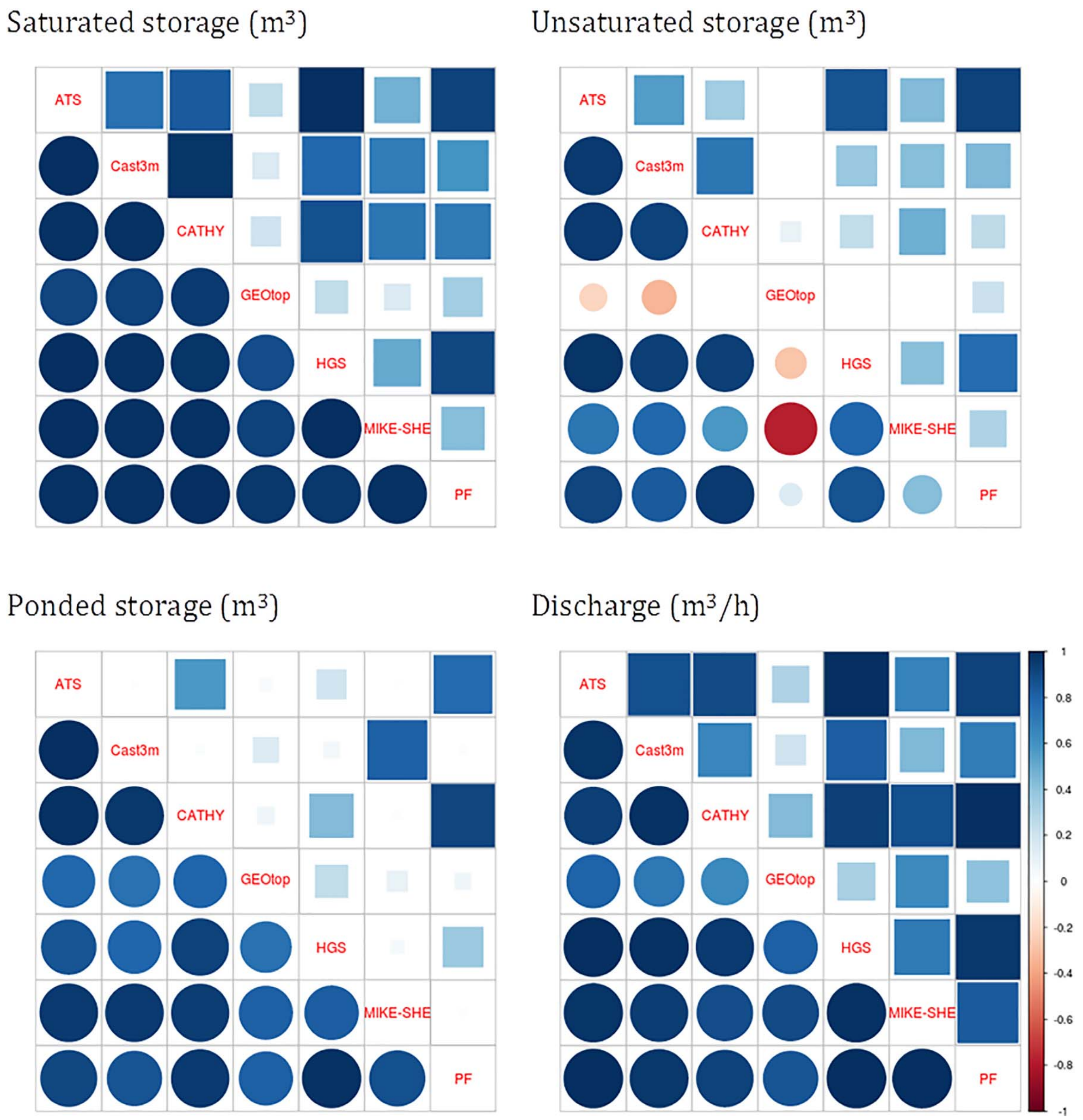

Figure 5. Results of the analysis of agreement for SC1, the tilted v-catchment benchmark. Pearson correlation values ($-1 \leq r \leq 1$) are plotted as circles below the diagonal. Duveiller biases ($0 \leq \alpha \leq 1$) are plotted as squares above the diagonal. Both, the size and color intensity of the circles and squares are proportional to the magnitude of the respective coefficient. White (blank) matrix entries mean $\alpha = 0$ or $r = 0$.

the agreement in unsaturated and saturated storages is good (large r and α values) except for MIKE-SHE, which exhibits the smallest saturated storage and largest unsaturated storage with different initial conditions compared to all other models (Figures 8 and 9). The difference in initial unsatured storage in MIKE-SHE could be due to an automatic adjustment of water contents in MIKE-SHE which occurs when the retention curve is too steep. The superslab case is based on van Genuchten parameters which results in a steep retention curve.

The ponding storage time series exhibits two periods of surface storage between 0 and 3 h and between about 6 h and 12 h simulation time. The first event is due to excess infiltration runoff generation along slab 2, which has a lower-saturated conductivity compared to the rainfall rate. Excess saturation ponding, i.e., the intersection of the perched water table with the land surface at slab 1, also contributes to the total ponded storage over this time period. A second, smaller event exists due to excess saturation ponding at the outlet. All models capture the different processes with some intermodel variability, reflected in r and α values close to 1 (Figure 9), which is acceptable, given the small magnitude of the events.

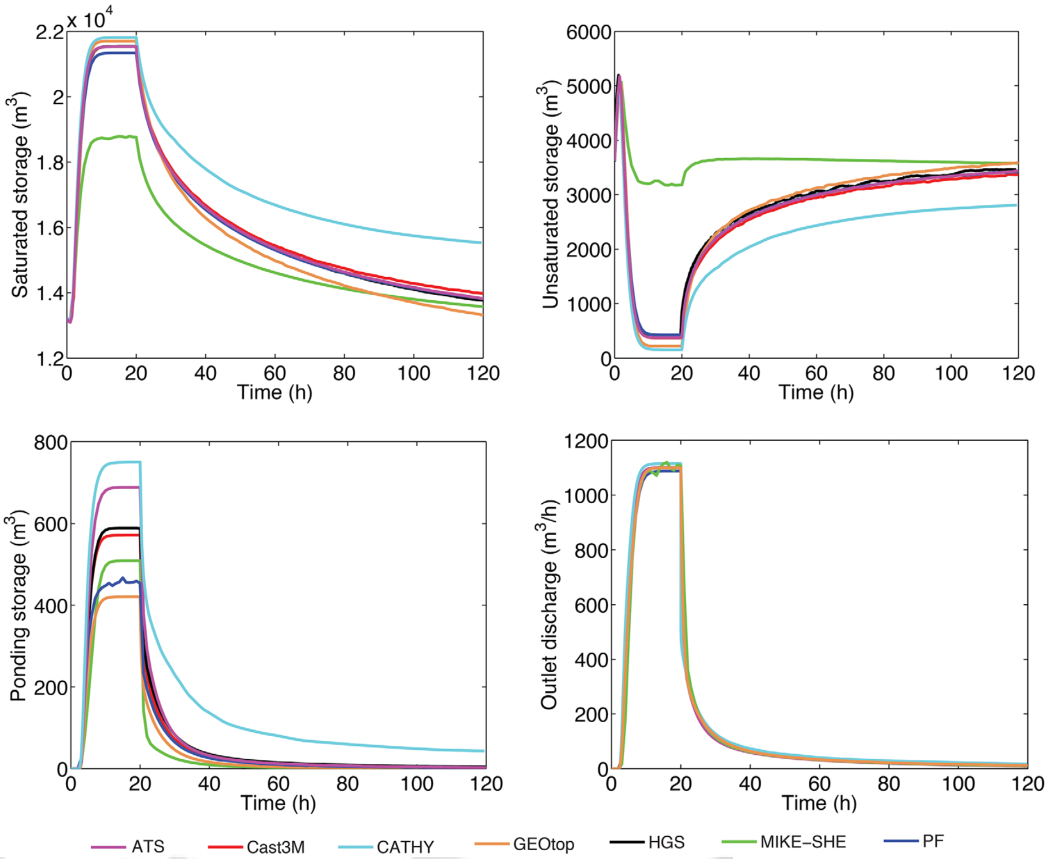

Figure 6. Storage and discharge time series of scenario S2 of the tilted v-catchment benchmark: saturated, unsaturated, ponded storage, and discharge at the outlet.

Discharge curves at the outlet and at slabs 1 and 2 show similar behavior. At the outlet, MIKE-SHE shows an early discharge peak due to runon from the slabs 1 and 2, which is not the case in the other models, where water infiltrated into the subsurface at the the ends of the slabs. The second discharge peak is due to equilibration of the profile toward hydrostatic conditions similar to scenario 1 of the tilted-v catchment. However, the pattern of the discharge curves is different because of the slightly different drainage history of the different models. HGS simulates very small discharge at early times and no discharge at later times and is therefore negatively correlated (Figure 9) with the other models, while GEOtop shows a higher discharge, similar to the of the tilted-v catchment case. Discharge at slab 1 shows the largest variability (small r and α values); however, absolute values are very small (essentially zero) and depend on details of the coupling and solver implementation. For example, in Cast3M runoff in the surface layer is simulated only if the water depth is greater than 10^{-10} m. Also, the relative water volume error in the Picard iteration is equal to 10^{-4}. The strict pressure continuity at the surface-subsurface interface represents a diffcult problem for the Picard algorithm, which often oscillates between two sets and fails to converge when the flow at the interface is small. Additionally, the wetness of the runoff layer (see section 2.1.2) may change from one time step to the next, which may lead to oscillations as well. For the discharge at slab 2, generated purely by excess infiltration, the curves agree well, which is reflected in r and α values close to 1 (Figure 9).

In Figure 9, an almost identical pattern of agreement as in Figure 7 can be identified for the subsurface and ponding storages, which shows generally high r and α values except for MIKE-SHE unsaturated storage. However, no distinct pattern of agreement emerges in the case of discharge at the outlet and at slab 1, when almost all model pairs show small correlations and α values. In contrast, all simulation results show high r and α values for the slab 2 discharge, suggesting that all models adequately model the process of pure infiltration excess runoff.

Figure 10 shows two cross sections of relative saturation S for each model at times 1.5 h (in the middle of the rain event) and 6 h (3 h into the recession). The cross sections accurately reflect the complexity in the

AGU Water Resources Research 10.1002/2016WR019191

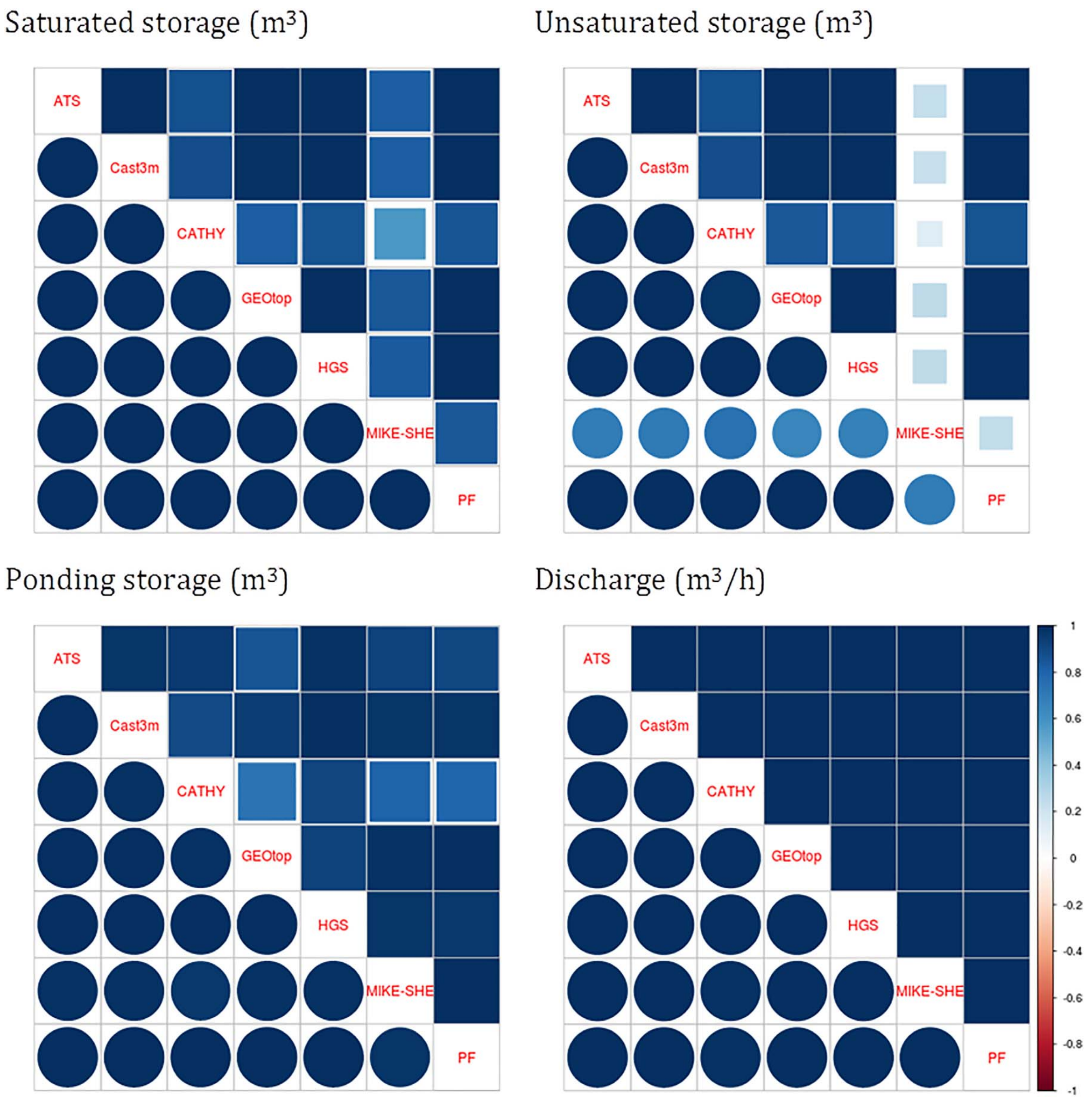

Figure 7. Results of the analysis of agreement for SC2, the tilted v-catchment benchmark. Pearson correlation values ($-1 \le r \le 1$) are plotted as circles below the diagonal and Duveiller biases ($0 \le \alpha \le 1$) are plotted as squares above the diagonal. Both, the size and color intensity of the circles and squares are proportional to the magnitude of the respective coefficient.

spatial distribution of S due to ponding along the low-conductivity slab in the center of the domain and the ensuing runon. The responses to this ponding and runon include preferential recharge, generation of a perched water table due to the horizontal low-conductivity slab, and recharge and equilibration of the fully saturated compartment, because of precipitation and gravity-driven drainage and lateral flow. In general, the continuum models agree reasonably well, with some differences in regions of preferential recharge and large conductivity contrasts, which also results in deviations in individual S profiles shown in Figure 11. F11 These differences are especially pronounced along infiltration fronts and close to the water table. The location of the water table is defined where saturation becomes $S < 1$ from one model layer to the next moving upward from the bottom of the domain. The discrepancies increase for MIKE-SHE owing to the coupling scheme for the saturated-unsatured zone, which apparently decouples the shallow from the deeper subsurface during the rainfall event. During the recession, all seven models start to converge, producing similar saturation distributions 3 h after cessation of rain. Some more distinct differences remain below slab 2 and in the water table depth, which also explains the different temporal onsets of outlet discharge at around 6 h of simulation time.

AGU Water Resources Research 10.1002/2016WR019191

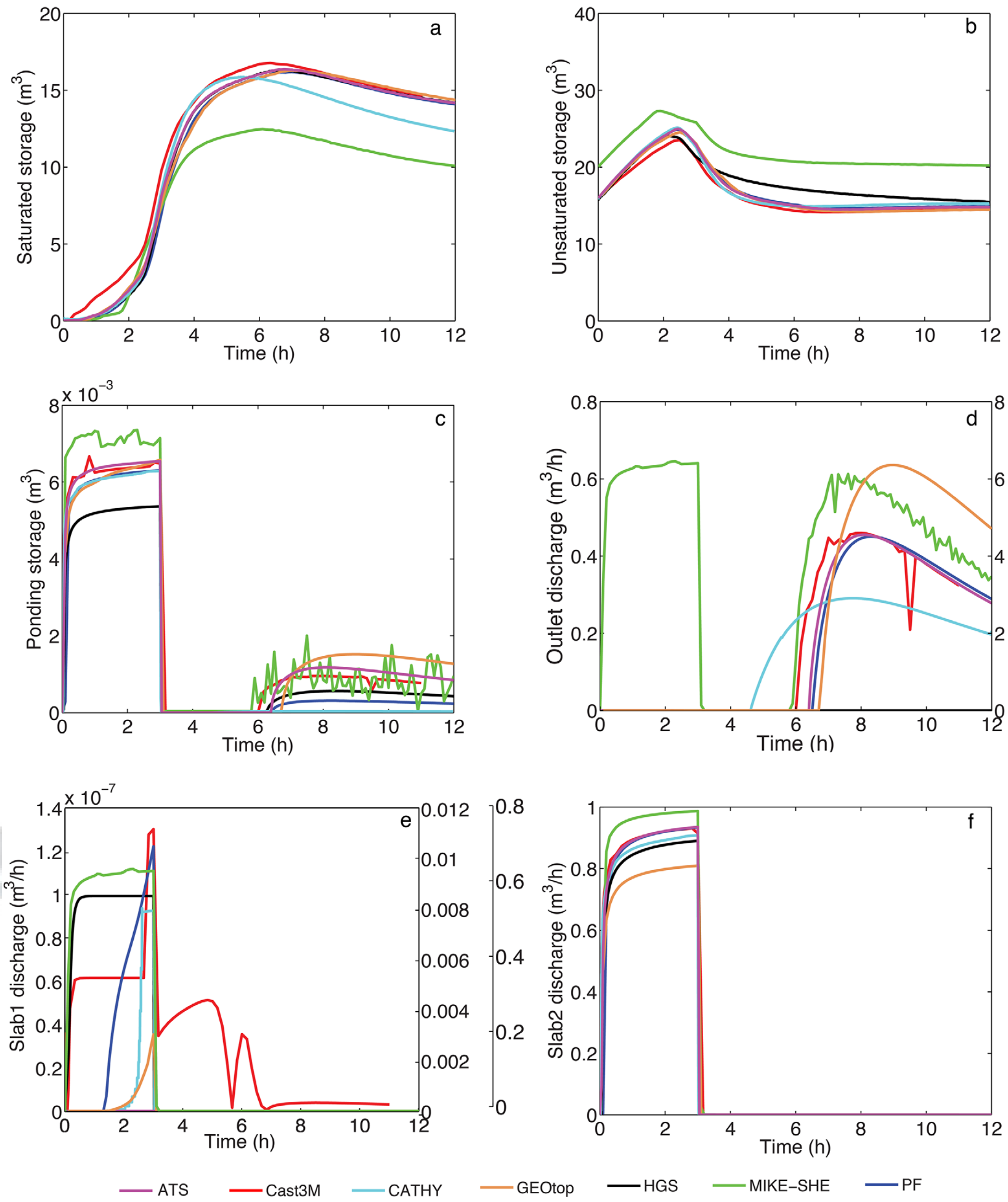

Figure 8. Storage and discharge time series of the superslab benchmark: (a) saturated storage, (b) unsaturated storage, (c) ponded storage, (d) hillslope discharge, (e) discharge at the horizontal slab (slab 1), and (f) discharge at the surface slab (slab 2). Note that for the outlet discharge GEOtop is plotted on the secondary axis, while for the slab 1 discharge Cast3M and HGS are plotted on the primary axis, CATHY, GEOtop, and PF are plotted on the secondary axis, and MIKE-SHE are plotted on the tertiary axis.

While Figure 10 provides some insight into the spatial variation of the results, Figure 11 shows discrete vertical saturation profiles at three different times (1.5, 3, and 6 h) of the simulations at three different locations (0, 8, and 40 m) along the hillslope coinciding with the outlet, slab 1, and slab 2, respectively. The different vertical discretizations used in the simulations are also evident from Figure 10 as well. Apparent oscillations visible along material boundaries (e.g., Cast3M, CATHY) are due to the discretization scheme (finite difference/control volume, terrain following). For example, in Cast3M, the grid cells are not horizontal but terrain following parallel to the surface slope. This tilted grid matches perfectly with slab 2 and the boundary conditions of the domain, while the discretization and ensuing visualizations creates artifacts in the case of the horizontal slab 1. At the outlet, the profiles agree well, but with MIKE-SHE deviating from the continuum models. More pronounced deviations between the location of the infiltration front computed by the models

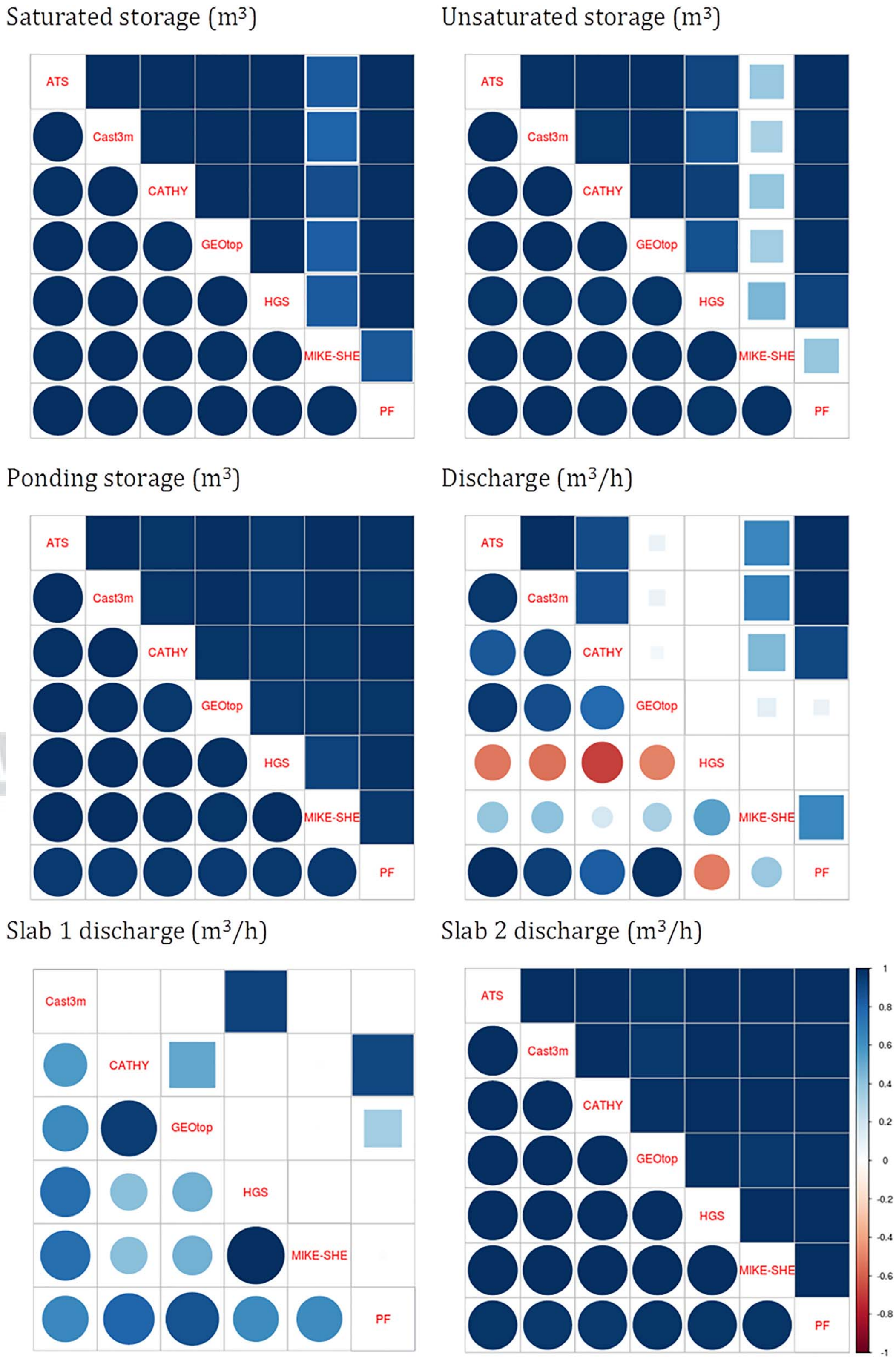

Figure 9. Results of the analysis of agreement for the superslab benchmark. Pearson correlation values ($-1 \leq r \leq 1$) are plotted as circles below the diagonal and Duveiller biases ($0 \leq \alpha \leq 1$) are plotted as squares above the diagonal. Both, the size and color intensity of the circles and squares are proportional to the respective coefficient. Missing matrix entries mean $\alpha = 0$. Note that ATS did not simulate discharge at slab 1, thus it is missing in the slab 1 discharge matrix.

AGU Water Resources Research 10.1002/2016WR019191

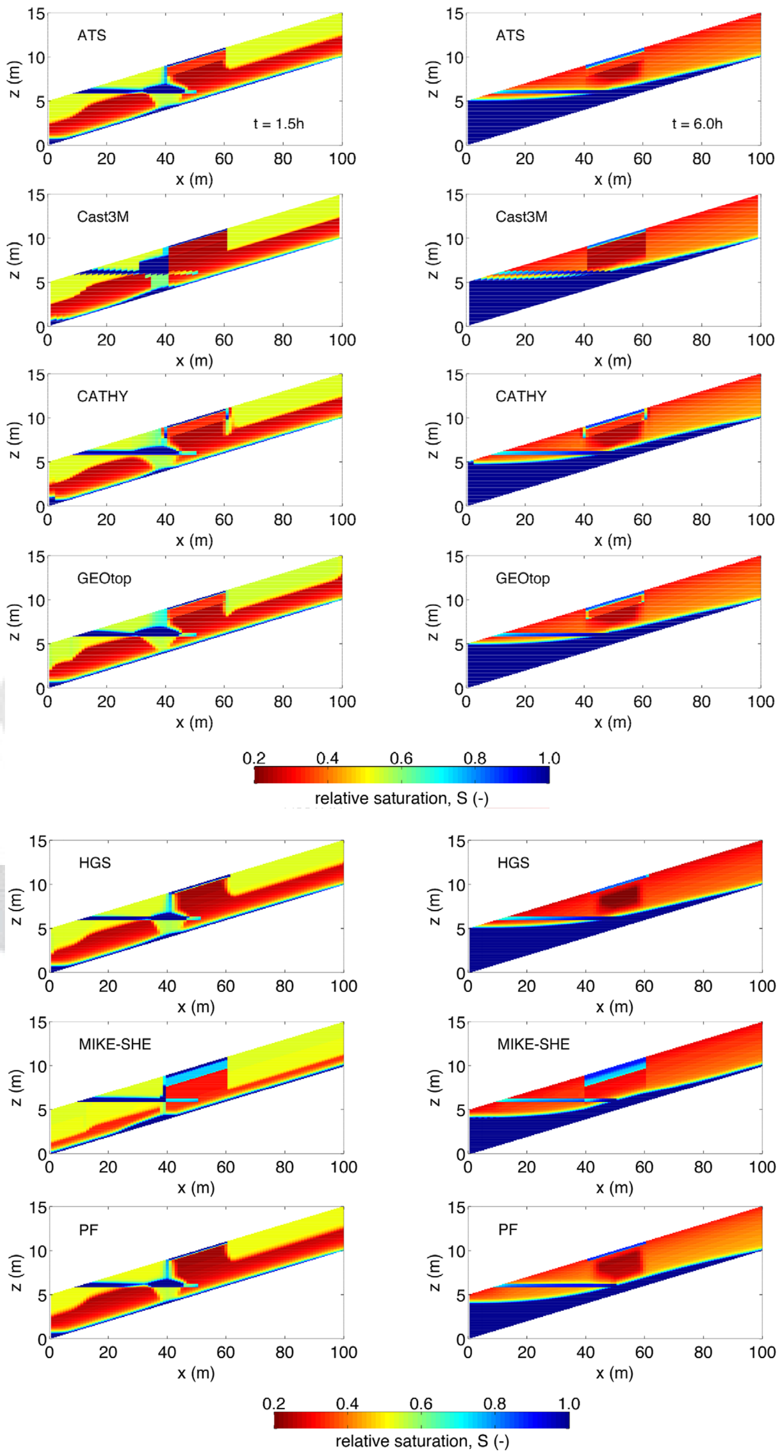

Figure 10. Cross sections of relative saturation S at (left column) $t = 1.5$ h and (right column) $t = 6$ h for the different models.

is observed for $x = 40$ m and times 1.5 and 3 h, where the heterogeneity of slab 2 and preferential infiltration due to runon impact the dynamics. Cast3M and MIKE-SHE simulate strong vertical saturation due to perched water on slab 1 extending close to the top of the land surface. In the other models, perched water is laterally distributed and infiltrates more efficiently into deeper parts of the profile. In general, it appears that MIKE-SHE underestimates lateral transport processes because of the one-dimensional vertical

AGU Water Resources Research 10.1002/2016WR019191

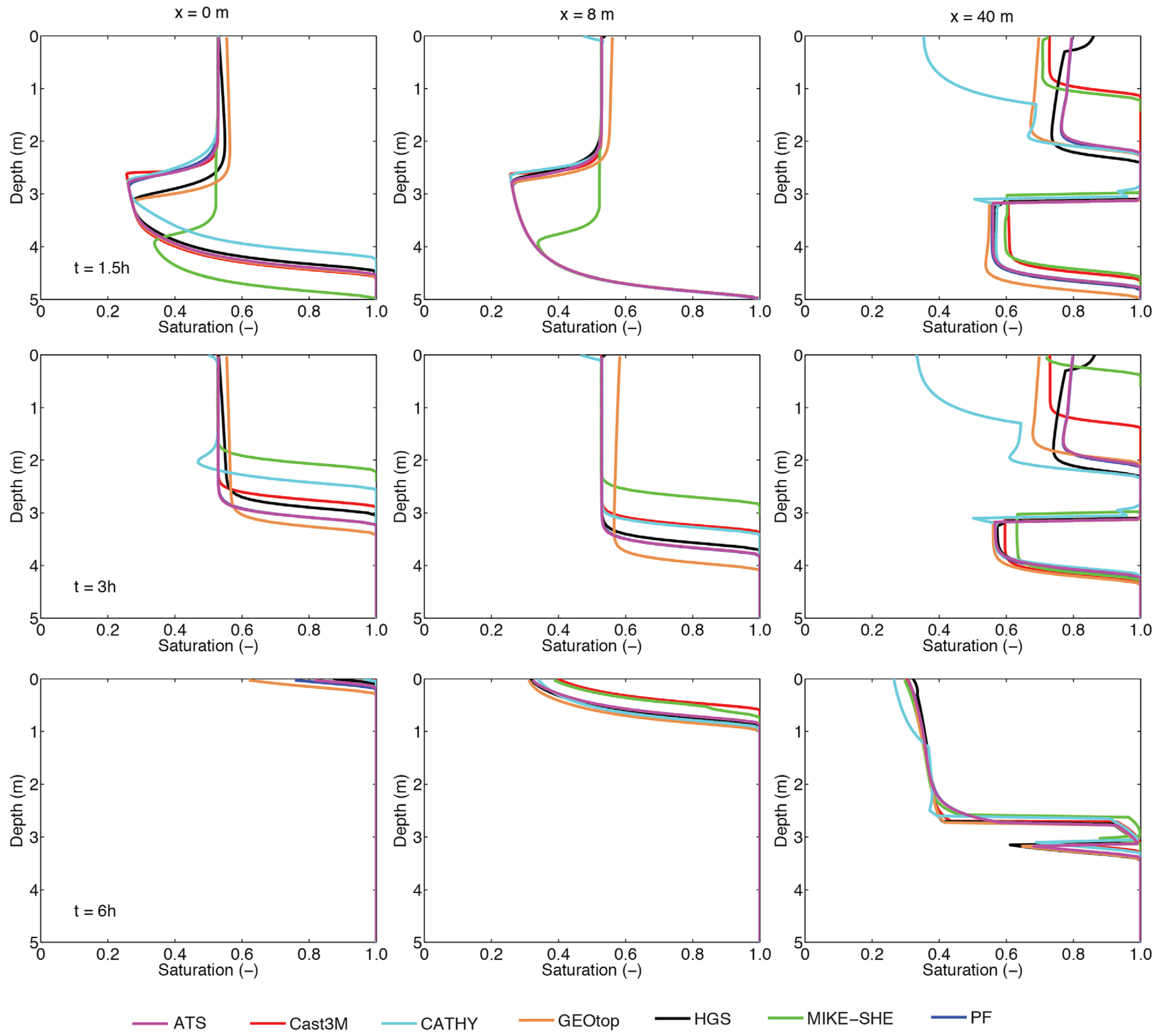

Figure. Profiles of relative saturation S at three different locations ($x = 0$, 8, and 40 m) along the z direction at time $t = 1.5$, 3, and 6 h.

discretization of the unsaturated zone. This intermodel uncertainty/inaccuracy should be taken into account in the comparison to observations by applying, for e.g., model ensembles and introducing model errors in the inversion and data assimilation algorithms. After 6 h of simulation time (3 h into the recession), all models agree remarkably well, reproducing a distinct peak in S at depth of around 3.2 m, which is due the the moisture remaining perched on slab 1.

5.3. Borden

The Borden benchmark reflects well the challenges in accurately simulating and reproducing discharge in a real-world setting. Note that the original topography was reinterpolated to accommodate the different discretization schemes used by the various models (finite difference/finite element/finite volume; structured/unstructured) and hence the results shown here are slightly different from the previously published results even with the same model [e.g., *Jones et al.*, 2006]. Because of the different discretization schemes used in the models (Table 1), the total model areas, and thus the storages, differ. Therefore, the storage estimates were normalized by the individual model areas, which are ATS = 975.25 m^2, Cast3m = 975.25 m^2, CATHY = 1022.25 m^2, GEOtop = 1022 m^2, HGS = 1022.25 m^2, MIKE-SHE = 1000 m^2, and PF = 1022.25 m^2. In the case of discharge and the comparison to the measured hydrograph from the original experiment, no

AGU Water Resources Research 10.1002/2016WR019191

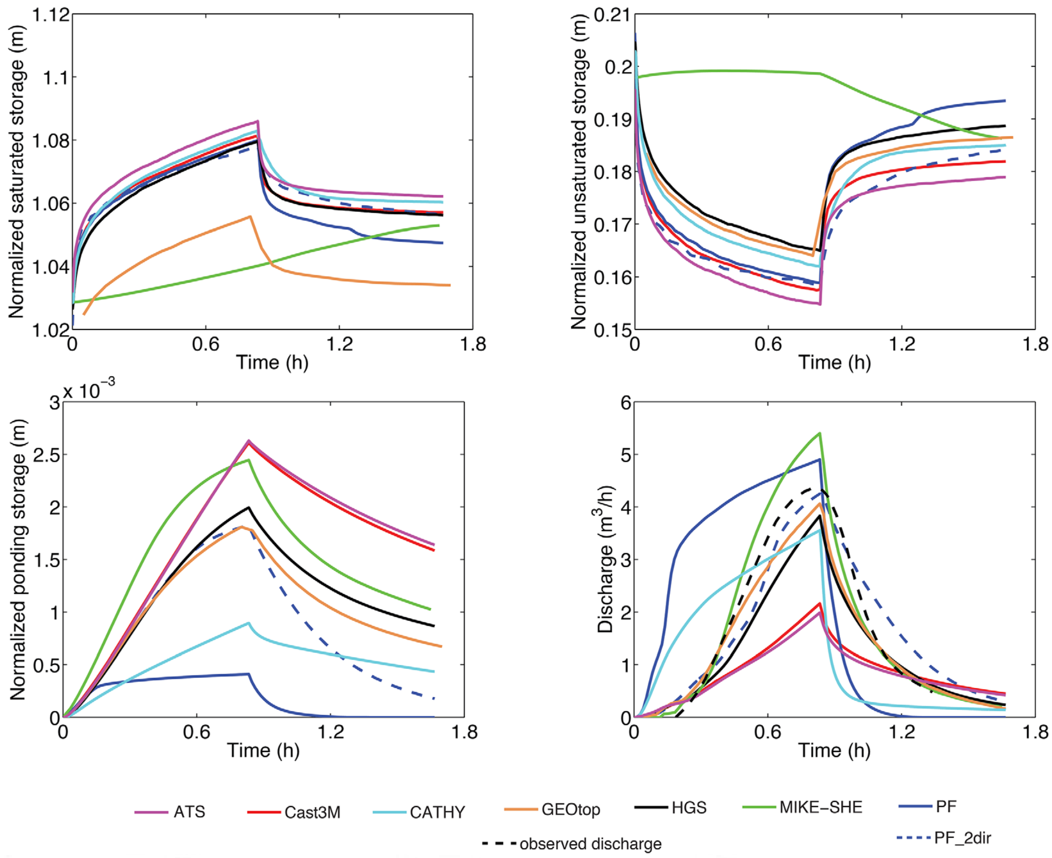

Figure 12. Storage and discharge time series of the Borden benchmark: (a) unsaturated storage, (b) saturated storage, (c) ponded storage, and (d) discharge. Note that storage values are normalized by the catchment area, which differs between the models because of different discretization schemes.

normalization was performed, because in the original study by *Abdul and Gillham* [1989] the area of the test site is only provided aproximately. In general, the continuum models arrive at quite consistent hydrologic responses with regard to the storages in terms of dynamics and absolute values, but some differences can be noted with respect to GEOtop (reduced α values) and more significantly MIKE-SHE (Figures 12 and 13). F12 F13 MIKE-SHE's subsurface storage response is essentially not correlated with the other models and cannot be explained satisfactorily at this point. Additional inspection of the numerical implementation would be needed. On the other hand, MIKE-SHE exhibits similar ponding storage dynamics and absolute values (high *r* and α values) to the other models in spite of having very different dynamics of unsaturated and saturated storages. This suggests a rather loose coupling between subsurface and surface water flow domains.

With respect to discharge at the outlet, ATS and Cast3M arrive at relatively small values also compared to the measurements from the original experiment by *Abdul and Gillham* [1989], which is due to the smaller catchment area of these models and thus less total water available for discharge from precipitation in these models. GEOtop, CATHY, and PF reproduce the peak discharge quite well; however, for the last two models, the rising limb is not well captured, which will be discussed in more detail below. MIKE-SHE reproduces the discharge hydrograph quite well (rising limb and recession), but slightly overestimates the peak discharge. In addition, MIKE-SHE exhibits similar ponded storage dynamics and absolute values compared to the other models, in spite of having very different dynamics of unsaturated and saturated storages: saturated storage is monotonically increasing, while unsaturated storage first increases during the rainfall and then decreases during the recession, contrary to the behavior of all the other models. Again the explanation could be the 1-D assumption in MIKE-SHE, which limits the increase in saturated storage to a small area at the bottom of the ditch, while in the hillslopes the infiltration in the variably saturated columns is only vertical and does not reach the groundwater table within the simulation period. In contrast, the continuum models generate horizontal unsaturated flow driven by the steep topography and leading to a faster saturation near the

AGU Water Resources Research 10.1002/2016WR019191

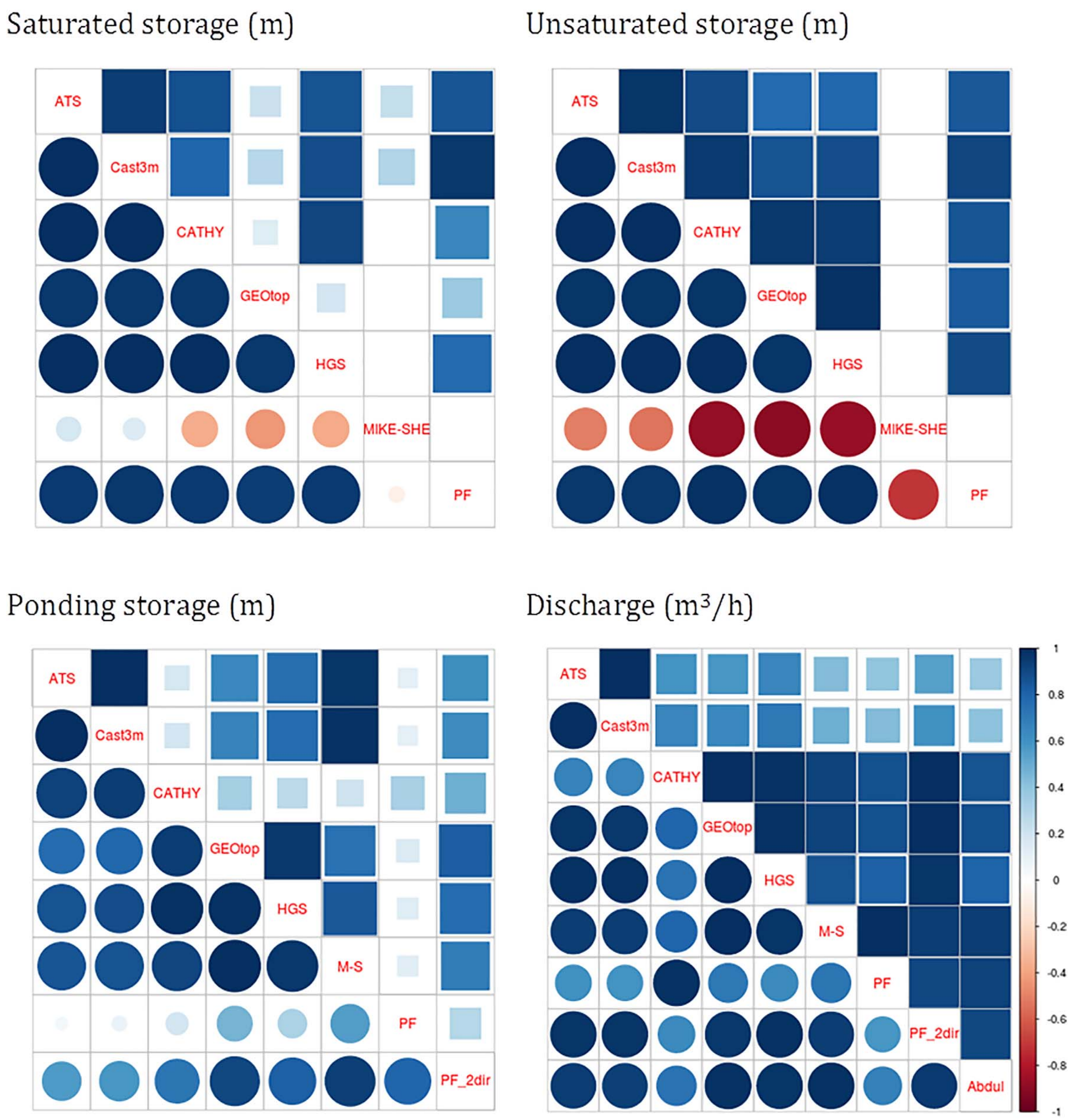

Figure 13. Results of the analysis of agreement for the Borden benchmark. In the case of ponded storage and discharge, results of MIKE-SHE and the two-directional PF simulation are indicated with the abbreviations M-S and PF_2dir, respectively. The Pearson correlation values ($-1 \leq r \leq 1$) are plotted as circles below the diagonal and Duveiller biases ($0 \leq \alpha \leq 1$) are plotted as squares above the diagonal. Both, the size and color intensity of the circles and squares are proportional to the respective coefficent. Missing matrix entries mean $\alpha = 0$.

bottom of the hillslope. The reason that Cast3M, HGS, and MIKE-SHE coincide closely with regard to ponded storage yet diverge significantly in the hydrograph response is again related to the different catchment areas that were used to normalize the storages but not the discharge.

It seems that the discharge is very sensitive to the elevation data and the derived topographic slopes used in the different models. In the Cast3M simulations, the mesh is generated from the 0.5 m resolution DEM, assuming that the raster values describe the nodal elevations of the cells. Hence, the simulated domain is smaller than for cell centered discretization schemes such as PF. As a consequence, all storages are smaller. The lack of additional surface storage and delayed runoff due to pits may contribute to the differences in discharge behavior in the case of CATHY and PF, which are based on the kinematic wave approximation and thus require the removal of any depression prior to the calculation of the friction slope. However, this does not explain why the other models exhibit ponding with only minor discharge at early simulation times,

AGU Water Resources Research 10.1002/2016WR019191

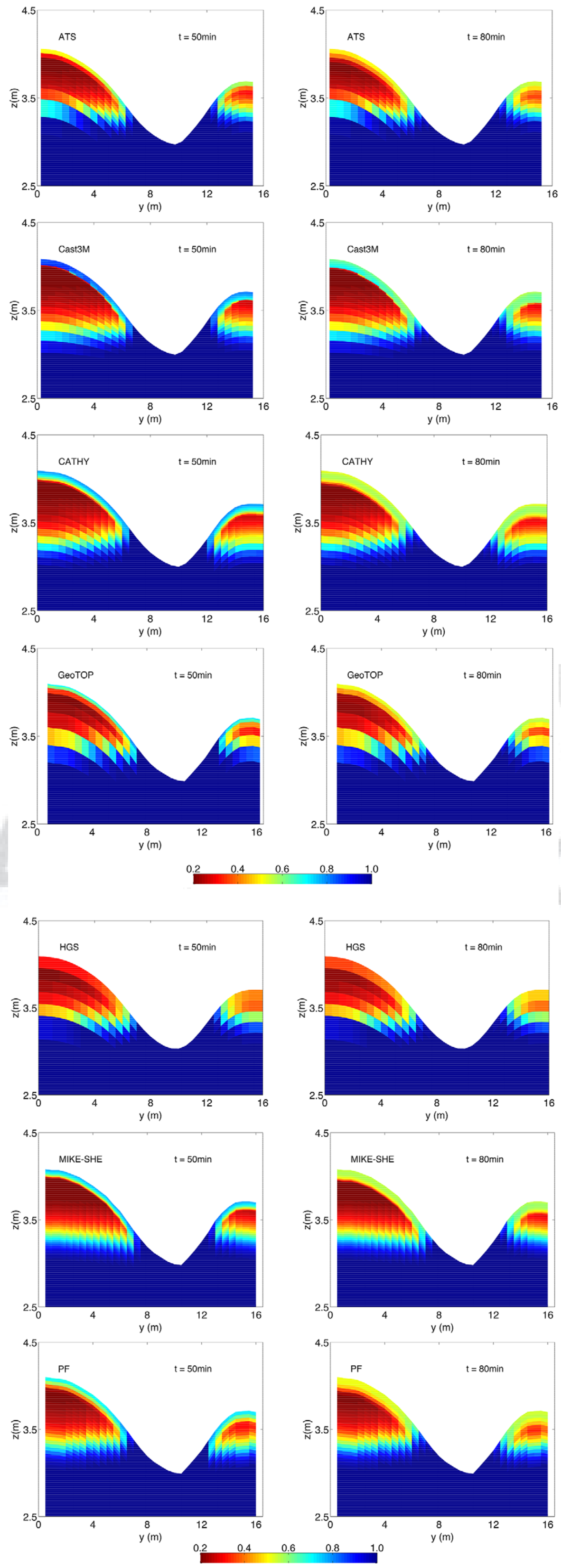

Figure 14. Profiles of relative saturation *S* in the center of the ditch perpendicular to the channel at times $t = 50$ and 80 min.

although the ditch outlet is indeed the lowest point in the model, which should thus produce instant discharge in case of ponding.

In order to interrogate the sensitivity to the topographic and friction slopes derived from the DEM, an additional PF simulation was performed, where friction slopes were calculated in both the x and y directions instead of unidirectional. The results of this additional simulation is called *PF_2dir* (two-directional) in Figures 12 and 13. The impact is remarkable, resulting in a completely different, more diffusive discharge behavior (and increased ponded storage) similar to the discharge responses calculated with diffusive wave models and close to the observed hydrograph (r and α values close to 1). Thus, the result suggests that in case of highly resolved DEMs, the method of friction slope estimation and topographic slope calculation may be as or more important than the approximation method used for the shallow water equation (kinematic versus diffusive wave).

In Figure 13, the results of the analysis of agreement show a pattern of consistency for the continuum models for the subsurface with a bias in the case of GEOtop-saturated storage. The strength of the pattern decreases for ponded storage, where PF and CATHY show a decrease in r and α values. Similar patterns emerge for the discharge, yet ATS and Cast3M now exhibit the smallest α values, because of a relative underestimation of discharge.

Figure 14 juxtaposes the different saturation, F14 *S*, profiles at times 50 and 80 min of the simulation. The lateral extent of saturation in the channel and the vertical *S* distribution at a certain distance from the land surface (approximately 5 to 10 cm) agrees well between the different models. The results suggest that lateral moisture transport in the unsaturated zone does not play a major role in the subsurface hydrodynamics; i.e., MIKE-SHE profiles do not deviate significantly from the other models. On the other hand, *S* values deviate significantly between the different models in the shallow subsurface close to the land surface, where the infiltration front is located. At the infiltration, front steep slopes and strong nonlinearity in the moisture-pressure and conductivity-pressure relationships exist, thus small differences in saturation between the

AGU Water Resources Research 10.1002/2016WR019191

different models may result in significantly different fluxes close to the land surface in the vicinity of the ditch. Again, this intermodel uncertainty/inacuracy should be taken into account in the comparison to and interpretation of observations in real-world applications, applying model ensembles and introducing model errors in inversion and data assimilation experiments.

6. Summary and Conclusions

Seven integrated hydrologic models were compared in three benchmark cases consisting of (i) a tilted v-catchment; (ii) a hillslope with two low-conductivity slab-heterogeneities (superslab case); and (iii) the Borden case, involving a shallow ditch on the order of 80 m in length. Each model was constructed based on the predefined input data with spatial and temporal resolutions based on the discretization scheme and computational capabilities of each individual model. Storage (saturated, unsaturated, ponding) and discharge dynamics, and, in addition for the superslab and Borden benchmarks, saturation cross sections and profiles, were analyzed in order to identify challenges in modeling the interactions of surface and subsurface water flow. An analysis of agreement that includes unsystematic and systematic deviations due to biases was performed between all models in a pairwise fashion.

Overall the models agree well in terms of temporal dynamics, yet exhibit differences in terms of absolute values, which is especially true for the models that treat the saturated, unsaturated, and also overland flow compartments as a single continuum. MIKE-SHE appears to generally have a lower level of agreement with the other models for the subsurface storages. The recession simulations of, for example, the v-catchment case (scenario S1) show that subtle dynamics are challenging to simulate, in particular with respect to ponded storage and absolute values, which may be relevant in inundation modeling. On the other hand, strong excitation by heavy rainfall results in quite uniform responses across all models, which lends confidence in the capabilities of the models to simulate hydrologic responses due to heavy rain events approaching or reaching steady state that are related to the processes of infiltration excess runoff and saturation. In some models, the representation of strong subsurface heterogeneities is a challenge and may result in deviations from a common modeling response between the different models, as in the case of the superslab. However, comparison of results along cross sections and local profiles shows good agreement between the models, which is remarkable. However, intermodel variability in local saturation values, especially along infiltration fronts and near the water table, should be taken into account in the comparison to in situ measurements. Comparison of the continuum models and MIKE-SHE suggests that the 1-D simplification in the unsaturated zone in MIKE-SHE's coupled compartment models may result in distinctly different storage dynamics and values, which is also partly due to the nature of the coupling of the saturated and unsaturated zones in the latter. The Borden benchmark demonstrates the challenge to arrive at consistent hydrologic rainfall responses in real-world settings, even in a quasi-laboratory setup and with only saturation excess runoff i.e., rather simple runoff generation dynamics. For example, differences in catchment area due to different discretization schemes lead clearly to differences in discharge. Thus, the results re-emphasize that special care must be taken in the setup of the model geometry. The PF example with one and two-directional friction slopes highlights the sensitivity of the hydrologic response with respect to discharge and internal storages and dynamics to calculations of overland flow. Multidirectional slopes lead to a more diffusive response due to more tortuous flow paths at the land surface. In particular, when the kinematic wave approximation is used, special attention must be paid to the derivation of the slopes. In the case of the diffusive wave approximation, the extra lateral diffusion alleviates this problem considerably. The presence or absence of an explicit channel and outlet, which must be derived from the DEM in a preprocessing step, might also play a relevant role in estimation of discharge.

It seems that the major difference between the simulated storages in the continuum models and MIKE-SHE originates from the 1-D assumption in the unsaturated zone in the latter, which is important in these small scale experimental setups with large topographical gradients (e.g., 7% for the Borden case). In order to analyze the implications of this simplification and the possible effect of the coupled model approach on the larger scale, a comparison of models at the river catchment scale would be of great interest and is being planned in future.

AGU Water Resources Research 10.1002/2016WR019191

Acknowledgments
We gratefully acknowledge the financial support of the IH-MIP2 workshop in Bonn in June 2013 by the German Research Foundation and the U.S. National Science Foundation, the first hydrology workshop with cofunding from both institutions to our knowledge. Additional sponsorship by the SFB/TR32, the Geoverbund ABC/J, and the Monalisa fund of the Province of Bolzano (Italy) is also gratefully acknowledged. We wish to thank M. Dall'Amico for help in performing the GEOtop simulations and the three anonymous reviewers for their constructive comments and suggestions. The data used in the studied can be obtained from the corresponding author.

References

Abbott, M. B., J. C. Bathurst, J. A. Cunge, P. E. Oconnell, and J. Rasmussen (1986), An introduction to the european hydrological system: Systeme hydrologique Europeen, She .2. Structure of a physically-based, distributed modeling system, *J. Hydrol.*, *87*(1–2), 61–77.

Abdul, A. S., and R. W. Gillham (1989), Field studies of the effects of the capillary-fringe on streamflow generation, *J. Hydrol.*, *112*(1–2), 1–18.

Aquanty, I. (2015), *HydroGeoSphere User Manual*, 435 pp., Waterloo, Ont. AQ3

Bixio, A. C., G. Gambolati, C. Paniconi, M. Putti, V. M. Shestopalov, V. N. Bublias, A. S. Bohuslavsky, N. B. Kasteltseva, and Y. F. Rudenko (2002), Modeling groundwater-surface water interactions including effects of morphogenetic depressions in the Chernobyl exclusion zone, *Environ. Geol.*, *42*(2–3), 162–177.

Bonetti, S., G. Manoli, J. C. Domec, M. Putti, M. Marani, and G. G. Katul (2015), The influence of water table depth and the free atmospheric state on convective rainfall predisposition, *Water Resour. Res.*, *51*, 2283–2297, *doi*:10.1002/2014WR016431.

Bowling, L. C., et al. (2003), Simulation of high-latitude hydrological processes in the Torne-Kalix basin: PILPS phase 2(e)—1: Experiment description and summary intercomparisons, *Global Planet Change*, *38*(1–2), 1–30.

Brezzi, F., K. Lipnikov, and V. Simoncini (2005), A family of mimetic finite difference methods on polygonal and polyhedral meshes, *Math. Model Meth. Appl. Sci.*, *15*(10), 1533–1551.

Brookfield, A. E., E. A. Sudicky, Y. J. Park, and B. Conant (2009), Thermal transport modelling in a fully integrated surface/subsurface framework, *Hydrol. Processes*, *23*(15), 2150–2164.

Butts, M. B., J. T. Payne, M. Kristensen, and H. Madsen (2004), An evaluation of the impact of model structure on hydrological modelling uncertainty for streamflow simulation, *J. Hydrol.*, *298*(1–4), 242–266.

Camporese, M., C. Paniconi, M. Putti, and S. Orlandini (2010), Surface-subsurface flow modeling with path-based runoff routing, boundary condition-based coupling, and assimilation of multisource observation data, *Water Resour. Res.*, *46*, W02512, *doi*:10.1029/2008WR007536.

Coon, E. T., J. D. Moulton, and S. L. Painter (2016), Managing complexity in simulations of land surface and near-surface processes, *Environ. Modell Software*, *78*, 134–149.

Dall'Amico, M., S. Endrizzi, S. Gruber, and R. Rigon (2011), A robust and energy-conserving model of freezing variably-saturated soil, *Cryosphere*, *5*(2), 469–484.

Davison, J. H., H. T. Hwang, E. A. Sudicky, and J. C. Lin (2015), Coupled atmospheric, land surface, and subsurface modeling: Exploring water and energy feedbacks in three-dimensions, *Adv. Water Resour.*, *86*, 73–85.

Della Chiesa, S., G. Bertoldi, G. Niedrist, N. Obojes, S. Endrizzi, J. D. Albertson, G. Wohlfahrt, L. Hortnagl, and U. Tappeiner (2014), Modelling changes in grassland hydrological cycling along an elevational gradient in the Alps, *Ecohydrology*, *7*(6), 1453–1473.

Duveiller, G., D. Fasbender, and M. Meroni (2016), Revisiting the concept of a symmetric index of agreement for continuous datasets, *Sci. Rep.*, *6*. AQ4

Ebel, B. A., B. B. Mirus, C. S. Heppner, J. E. VanderKwaak, and K. Loague (2009), First-order exchange coefficient coupling for simulating surface water-groundwater interactions: Parameter sensitivity and consistency with a physics-based approach, *Hydrol. Processes*, *23*(13), 1949–1959.

Endrizzi, S., S. Gruber, M. Dall'Amico, and R. Rigon (2014), GEOtop 2.0: Simulating the combined energy and water balance at and below the land surface accounting for soil freezing, snow cover and terrain effects, *Geosci. Model Dev.*, *7*(6), 2831–2857.

Gasper, F., K. Goergen, P. Shrestha, M. Sulis, J. Rihani, M. Geimer, and S. Kollet (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), *Geosci. Model Dev.*, *7*(5), 2531–2543.

Gottardi, G., and M. Venutelli (1993), A control-volume finite-element model for 2-dimensional overland-flow, *Adv. Water Resour.*, *16*(5), 277–284.

Graham, D. N., and M. B. Butts (2005), Flexible, integrated watershed modelling with MIKE SHE, in *Watershed Models*, edited by V. P. Singh and D. K. Frevert, pp. 245–272, Taylor and Francis, Boca Raton, Fla.

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

Jones, J. E., and C. S. Woodward (2001), Newton-Krylov-multigrid solvers for large-scale, highly heterogeneous, variably saturated flow problems, *Adv. Water Resour.*, *24*(7), 763–774.

Jones, J. P., E. A. Sudicky, A. E. Brookfield, and Y. J. Park (2006), An assessment of the tracer-based approach to quantifying groundwater contributions to streamflow, *Water Resour. Res.*, *42*, W02407, doi:10.1029/2005WR004130.

Kollet, S. J., and R. M. Maxwell (2006), Integrated surface-groundwater flow modeling: A free-surface overland flow boundary condition in a parallel groundwater flow model, *Adv. Water Resour.*, *29*(7), 945–958.

Kollet, S. J., and R. M. Maxwell (2008), Capturing the influence of groundwater dynamics on land surface processes using an integrated, distributed watershed model, *Water Resour. Res.*, *44*, W02402, doi:10.1029/2007WR006004.

Kollet, S. J., I. Cvijanovic, D. Schuttemeyer, R. M. Maxwell, A. F. Moene, and P. Bayer (2009), The Influence of Rain Sensible Heat and Subsurface Energy Transport on the Energy Balance at the Land Surface, *Vadose Zone J.*, *8*(4), 846–857.

Kollet, S. J., R. M. Maxwell, C. S. Woodward, S. Smith, J. Vanderborght, H. Vereecken, and C. Simmer (2010), Proof of concept of regional scale hydrologic simulations at hydrologic resolution utilizing massively parallel computer resources, *Water Resour. Res.*, *46*, W04201, doi:10.1029/2009WR008730.

Kurtz, W., G. W. He, S. J. Kollet, R. M. Maxwell, H. Vereecken, and H. J. H. Franssen (2016), TerrSysMP-PDAF version 1.0): A modular high-performance data assimilation framework for an integrated land surface-subsurface model, *Geosci. Model Dev.*, *9*(4), 1341–1360.

Larsen, M. A. D., J. H. Christensen, M. Drews, M. B. Butts, and J. C. Refsgaard (2016), Local control on precipitation in a fully coupled climate-hydrology model, *Sci. Rep.*, *6*.

Liggett, J. E., A. D. Werner, and C. T. Simmons (2012), Influence of the first-order exchange coefficient on simulation of coupled surface-subsurface flow, *J. Hydrol.*, *414*, 503–515.

Manoli, G., S. Bonetti, J. C. Domec, M. Putti, G. Katul, and M. Marani (2014), Tree root systems competing for soil moisture in a 3D soil-plant model, *Adv. Water Resour.*, *66*, 32–42.

Manoli, G., M. Rossi, D. Pasetto, R. Deiana, S. Ferraris, G. Cassiani, and M. Putti (2015), An iterative particle filter approach for coupled hydro-geophysical inversion of a controlled infiltration experiment, *J. Comput. Phys.*, *283*, 37–51.

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

Maxwell, R. M., and S. J. Kollet (2008), Quantifying the effects of three-dimensional subsurface heterogeneity on Hortonian runoff processes using a coupled numerical, stochastic approach, *Adv. Water Resour.*, *31*(5), 807–817.

Maxwell, R. M., and N. L. Miller (2005), Development of a coupled land surface and groundwater model, *J. Hydrometeorol.*, *6*(3), 233–247.

Maxwell, R. M., F. K. Chow, and S. J. Kollet (2007), The groundwater-land-surface-atmosphere connection: Soil moisture effects on the atmospheric boundary layer in fully-coupled simulations, *Adv. Water Resour.*, *30*(12), 2447–2466.

Maxwell, R. M., J. K. Lundquist, J. D. Mirocha, S. G. Smith, C. S. Woodward, and A. F. B. Tompson (2011), Development of a coupled groundwater-atmosphere model, *Mon. Weather Rev.*, *139*(1), 96–116.

Maxwell, R. M., et al. (2014), Surface-subsurface model intercomparison: A first set of benchmark results to diagnose integrated hydrology and feedbacks, *Water Resour. Res.*, *50*, 1531–1549, doi:10.1002/2013WR013725.

Maxwell, R. M., L. E. Condon, and S. J. Kollet (2015), A high-resolution simulation of groundwater and surface water over most of the continental US with the integrated hydrologic model ParFlow v3, *Geosci. Model Dev.*, *8*(3), 923–937.

Mielke, P. W. (1991), The application of multivariate permutation methods based on distance functions in the earth-sciences, *Earth Sci. Rev.*, *31*(1), 55–71.

Niu, G. Y., C. Paniconi, P. A. Troch, R. L. Scott, M. Durcik, X. B. Zeng, T. Huxman, and D. C. Goodrich (2014a), An integrated modelling framework of catchment- scale ecohydrological processes: 1. Model description and tests over an energy-limited watershed, *Ecohydrology*, *7*(2), 427–439.

Niu, G. Y., P. A. Troch, C. Paniconi, R. L. Scott, M. Durcik, X. B. Zeng, T. Huxman, D. Goodrich, and J. Pelletier (2014b), An integrated modelling framework of catchment- scale ecohydrological processes: 2. The role of water subsidy by overland flow on vegetation dynamics in a semi- arid catchment, *Ecohydrology*, *7*(2), 815–827.

Orlandini, S., and R. Rosso (1998), Parameterization of stream channel geometry in the distributed modeling of catchment dynamics, *Water Resour. Res.*, *34*(8), 1971–1985.

Osei-Kuffuor, D., R. M. Maxwell, and C. S. Woodward (2014), Improved numerical solvers for implicit coupling of subsurface and overland flow, *Adv. Water Resour.*, *74*, 185–195.

Panday, S., and P. S. Huyakorn (2004), A fully coupled physically-based spatially-distributed model for evaluating surface/subsurface flow, *Adv. Water Resour.*, *27*(4), 361–382.

Refsgaard, J. C., and B. Storm (1995), MIKE SHE, in *Computer Models of Watershed Hydrology*, edited by V. P. Singh, pp. p809–846, Water Resour. Publ., Colo. AQ5

Rigon, R., G. Bertoldi, and T. M. Over (2006), GEOtop: A distributed hydrological model with coupled water and energy budgets, *J. Hydrometeorol.*, *7*(3), 371–388.

Rossi, M., G. Manoli, D. Pasetto, R. Deiana, S. Ferraris, C. Strobbia, M. Putti, and G. Cassiani (2015), Coupled inverse modeling of a controlled irrigation experiment using multiple hydro-geophysical data, *Adv. Water Resour.*, *82*, 150–165.

Shrestha, P., M. Sulis, M. Masbou, S. Kollet, and C. Simmer (2014), A scale-consistent terrestrial systems modeling platform based on COSMO, CLM, and ParFlow, *Mon. Weather Rev.*, *142*(9), 3466–3483.

Smith, M. B., and H. V. Gupta (2012), The Distributed Model Intercomparison Project (DMIP): Phase 2 experiments in the Oklahoma Region, USA Preface, *J. Hydrol.*, *418*, 1–2.

Steefel, C. I., et al. (2015), Reactive transport codes for subsurface environmental simulation, *Comput. Geosci.*, *19*(3), 445–478.

Storm, B. (1991), Modelling of saturated flow and the coupling of surface and subsurface flow, in *Recent Advances in the Modelling of Hydrologic Systems*, edited by D. S. Bowles and P. E. O'Connel, pp. 185–203, Kluwer Acad., Netherlands.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl (2012), An overview of Cmip5 and the experiment design, *Bull. Am. Meteorol. Soc.*, *93*(4), 485–498.

van Genuchten, M. T. (1980), A closed-form equation for predicting the hydraulic conductivity of unsaturated soils, *Soil Sci. Soc. Am. J.*, *44*(5), 892–898.

Weill, S., E. Mouche, and J. Patin (2009), A generalized Richards equation for surface/subsurface flow modelling, *J. Hydrol.*, *366*(1–4), 9–20.

Wijesekara, G. N., B. Farjad, A. Gupta, Y. Qiao, P. Delaney, and D. J. Marceau (2014), A comprehensive land-use/hydrological modeling system for scenario simulations in the Elbow River Watershed, Alberta, Canada, *Environ. Manage.*, *53*(2), 357–381.