Virtual Experiments Guide Calibration Strategies for a Real-World Watershed Application of Coupled Surface-Subsurface Modeling

Xuan Yu[1]; Christopher Duffy[2]; Yu Zhang[3]; Gopal Bhatt[4]; and Yuning Shi[5]

Abstract: Virtual experiments have been designed for the development and validation of coupled surface-subsurface modeling. Potentially, virtual experiments can guide model calibration as well. To address the role of virtual experiments in model calibration, a novel approach was described for a real watershed calibration of Penn State Integrated Hydrologic Model (PIHM) guided by the V-shaped catchment simulation. First, a benchmarking experiment of coupled surface-subsurface modeling was developed and documented on the V-shaped catchment. Then, the performance of hydrologic predictions for the V-shaped catchment was calculated and demonstrated different levels of correlations. The correlations were found stable, which had the potential to be used as the weights of multiobjective calibration. Therefore, a weighted multiobjective calibration was developed for a real-world watershed by transferring the correlations obtained from the virtual experiments. Expectedly, the parameters calibrated using the weighted approach indicated improvement of the model performance in simulating water-table depths and evapotranspiration with little sacrifice of model performance in streamflow. In addition, this study also compares the weighted average calibration and unweighted calibration. The results demonstrate the weighted objective optimization achieved satisfactory compromise for each calibration objective. Overall, the virtual experiment is proved to be an efficient tool to facilitate calibration of complex models. The proposed weighted objective approach provides an effective calibration strategy for the multiple observation constraints, which can be applied for the calibration of coupled environmental process models with multiple observations. **DOI: 10.1061/(ASCE)HE.1943-5584.0001431.**

Author keywords: Calibration; Coupled surface-subsurface modeling; V-shaped catchment; Penn State Integrated Hydrologic Model (PIHM); Weighted objective function.

Introduction

Virtual experiment validation has been widely applied in the development of hydrologic models. Unlike real-world watershed applications, virtual experiments rely on artificial model domains [e.g., regular shapes (Henry 1964), simplified stream channels (Di Giammarco et al. 1996)] and simplified hydrological processes, so that the key hydrologic dynamics are captured with minimum tunable parameters (Weiler and McDonnell 2004). Usually, these virtual simulations can be used to derive analytical or semi-analytical solutions to guide the development and verification of hydrologic models (Di Giammarco et al. 1996; Simpson and Clement 2003; Zidane et al. 2012; Maxwell et al. 2014) because it can be easily reproduced and reused. Another important application of virtual experiment is synthetic modeling of hydrological responses in different conditions. For example, Di Giammarco et al. (1996) validated a mixed one- and two-dimensional approach of overland flow by comparing model results against analytical solution of the V-shaped schematic catchment. Miller (1995) analyzed topographic sensitivity of surface heat and moisture flux on a synthetic terrain. Mallard et al. (2014) used synthetic watersheds with varying topographic structure and stream network geometry to understand the geometry influence on stream water composition. Seo and Schmidt (2013) explored the relationship between channel network configurations and hydrograph sensitivity to storm kinematics on a series of synthetic circular catchments. Maxwell et al. (2014) presented the model results of a set of different integrated hydrologic models on standardized benchmark problems to understand the representation of coupled hydrologic processes. These modeling scenarios of virtual experiments improved the authors' understanding of the complexity of coupled hydrological processes. Predictably, virtual experiments could be a potential guidance of calibration process, which is rare in the literature.

One major challenge on the calibration of coupled surface-subsurface modeling is the computational cost. The mathematical formulation of the multiscale multiprocess physical system can be quite complicated, leading to a time-consuming numerical model (Kumar and Duffy 2015; Osei-Kuffuor et al. 2014). Calibration processes usually require a large number of model simulation runs to find the optimal parameter set. The calibration of such coupled surface-subsurface models not only causes exhaustive burden on

[1]Postdoctoral Researcher, Dept. of Geological Sciences, Univ. of Delaware, Newark, DE 19716; formerly, Ph.D. Student, Dept. of Civil and Environmental Engineering, Pennsylvania State Univ., University Park, PA 16802. E-mail: xuan@udel.edu

[2]Professor, Dept. of Civil and Environmental Engineering, Pennsylvania State Univ., University Park, PA 16802 (corresponding author). E-mail: cxd11@psu.edu

[3]Ph.D. Student, Dept. of Civil and Environmental Engineering, Pennsylvania State Univ., University Park, PA 16802. E-mail: yzz130@psu.edu

[4]Research Associate, Dept. of Civil and Environmental Engineering, Pennsylvania State Univ., University Park, PA 16802. E-mail: gxb913@psu.edu

[5]Research Associate, Dept. of Ecosystem Science and Management, Pennsylvania State Univ., University Park, PA 16802. E-mail: yshi@psu.edu

Note. This manuscript was submitted on August 18, 2015; approved on April 29, 2016; published online on July 19, 2016. Discussion period open until December 19, 2016; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Hydrologic Engineering***, © ASCE, ISSN 1084-0699.

the modelers but deters the attention from the purpose of hydrologic modeling. One solution is to utilize the statistically efficient optimization algorithms to search the optimal parameter set (Nicklow et al. 2010), which can decrease the number of model runs and save the total computational cost. Another solution is to minimize the calibration period (Juston et al. 2009; Singh and Bárdossy 2012). It is always worth finding the appropriate length of observation data for effective calibration of hydrologic models (Razavi and Tolson 2013; Vrugt et al. 2006). In general, a virtual experiment usually involves simplified hydrological processes with less computational cost, which could be an ideal test bed to analyze sensitivity of model parameters and to discover the effective calibration strategies.

Another challenge are the multistate objectives of coupled surface-subsurface models. In real-world applications, coupled surface-subsurface models try to capture the heterogeneity of hydrologic state and flux variables. Consequently, in order to understand other catchment behaviors beyond streamflow, e.g., surface-subsurface flow interactions and watershed connectivity between riparian zone and upslope area, it is desirable to calibrate and validate model predictions against multiple observed hydrologic states and fluxes in the watershed as a multiobjective problem. Under ideal conditions, these objectives should be nonconflicting, which means that one set of optimal parameters should be selected to satisfy all the objectives including different hydrologic states and fluxes to the maximum degree possible. In practical applications, due to errors in model structure (Butts et al. 2004), uncertainty in parameters (Shi et al. 2014), and observed data (McMillan et al. 2012), the objectives of fitting different hydrological responses are usually not satisfied simultaneously. Traditionally, optimization problems involving multiple and conflicting objectives have been solved by combining the objectives into a scalar function (aggregation function), and next solving the equivalent single-optimization problem to identify the best-compromise solution (Efstratiadis and Koutsoyiannis 2010; Hsie et al. 2014). Such aggregated multiobjective calibration studies tried to find the best weighting coefficients (Rozos et al. 2004; Li et al. 2010; Dung et al. 2011). It was found that a direct incorporation of these metrics of different measurements into the objective function might not be theoretically reasonable (Rozos et al. 2004). It is has been found that different optimal parameter sets can be obtained by changing the weighting coefficients. Rozos et al. (2004) followed a hybrid strategy based on a combination of automatic and manual methods by adjusting the weights according to previous optimization results. Dung et al. (2011) calibrated a hydrodynamic model against multiple stream water elevations. The weight of model performance at each gauging station was assigned according to the inundation impact. The comparison result showed that the unweighted calibration performed worse than the result of innudation-based weighted calibration. However, expert knowledge was required to subjectively assign the innudation-based weights (Dung et al. 2011). Therefore, the selection of weights and formulation of a weighted objective function may lead to controversial solutions according to the experiences of the modelers (Rientjes et al. 2013).

The aim of this paper is to describe and document the simulation of a virtual watershed, to explore the model performance correlation of different variable, and then to test if the performance correlation can be applied on the calibration of a real-world watershed. Specifically, the synthetic virtual experiment was developed to simulate the hydrologic process at the V-shaped catchment with the Penn State Integrated Hydrologic Model (PIHM). Then one rainfall runoff event was tested to justify if it contains sufficient information for the calibration on the V-shaped catchment. Based on the correlation of the PIHM performance at each observed variable, "informativeness" was developed. Finally, a weighted objective function was formulated according to the informativeness and was tested on the calibration of a real-world watershed.

Method

PIHM Description

PIHM is a physics-based, spatially distributed hydrologic model with a coupled surface-subsurface approach. It simulates the watershed processes including interception, throughfall, infiltration, recharge, evapotranspiration, overland flow, unsaturated soil water, groundwater flow, and channel routing in a fully coupled scheme (Qu and Duffy 2007). Evapotranspiration is calculated using the Penman-Monteith approach adapted from the Noah land surface model (Chen and Dudhia 2001). Overland flow is described in two-dimensional (2D) estimation of Saint-Venant equations. Movement of moisture in unsaturated zones is assumed to be vertical and is modeled using Richard's equation. The model assumes that each subsurface layer can have both unsaturated and saturated storage components. Balance equations of the unsaturated and saturated zones are formed in a fully coupled way, and the unsaturated conductivity is determined by the van Genuchten (1980) approach. The channel routing is modeled using one-dimensional (1D) estimation of Saint-Venant equations. PIHM uses diffusive wave approximation for channel routing and overland flow. For saturated groundwater flow, the 2D Dupuit approximation is applied (Qu and Duffy 2007). Spatially, the modeling domain is decomposed into Delaunay triangles (DTs). DT is an unstructured mesh that consists of a set P of points in a plane such that no point in P is inside the circumcircle of any triangle. PIHM allows users to resolve spatial data over the watershed and the DT can be constrained by point or vector data (e.g., stream gauge, wells, soil maps, and land cover) and the watershed boundary conditions (Kumar 2009). The model resolves hydrological processes for overland flow, channel routing, and subsurface flow, governed by a partial differential equation (PDE) system. The system is discretized on the triangular mesh and the projected prism from canopy to bedrock. PIHM uses a semidiscrete, finite-volume formulation for solving the system of coupled PDEs, resulting in a system of ordinary differential equations (ODEs) representing all processes within the prismatic control volume. Detailed descriptions of the rest of modeling theory and the full set of mathematical formulation can be found at PIHM (2016) and associated publications (Kumar 2009; Qu and Duffy 2007).

Virtual Experiment on the V-Shaped Catchment

The original V-shaped catchment was extended to represent the coupled surface and subsurface processes by incorporating a 2-m-deep subsurface domain (Fig. 1). The spatial mesh was generated by DistMesh (Persson and Strang 2004). The mesh of triangles was selected as coarse as possible so that the model computational cost is small. A sandy soil, Bordenan sand, was selected from unsaturated soil hydraulic database (UNSODA) (Leij 1996) to represent the homogenous subsurface layer. The subsurface parameters are listed in Table 1. The initial condition was prescribed as 1 m deep of saturated zone uniformly. The meteorological data of 2009 at Shale Hills was selected as the model forcing. To simplify real-world rainfall-runoff processes, only three rainfall equal rainfall (i.e., 38.1 mm/h for 6 h) events were applied to simulation of the hydrological responses. The idea of artificial rainfall events followed the irrigation experiment at Shale Hills in 1974 (Lynch 1976). Such a synthetic experiment with artificial rainfall and regular geometry showed simple hydrologic responses for single and

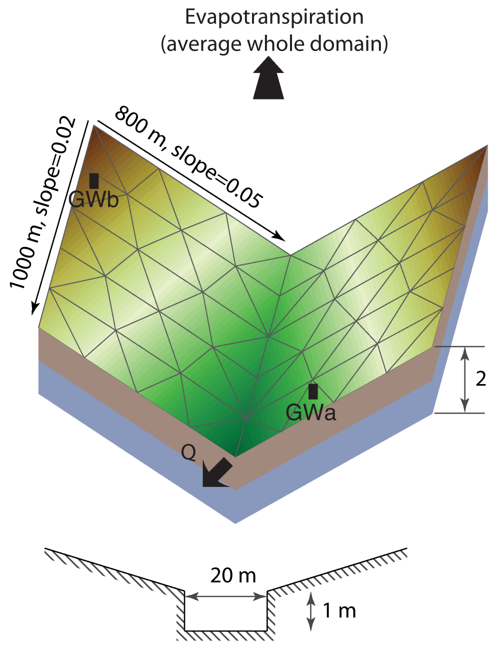

Fig. 1. Mesh of the V-shaped catchment and the observation locations

continual rainfall events (Fig. 2). The extended V-shaped catchment simulation was created as a benchmarking dataset, which is publicly accessible in Yu (2015).

Monte Carlo sampling was conducted throughout the feasible ranges (Yu et al. 2013) of each parameter (Table 2). The CPU time to obtain the 3-month synthetic simulation on a PC (processor 3.16 GHz, 4.0 GB RAM) was <3 min. There were 10,000 PIHM simulations in total. The Nash-Sutcliffe coefficient of efficiency (NSE) was calculated to evaluate model performances

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{t}(O_i - P_i)^2}{\sum_{i=1}^{t}(O_i - \bar{O})^2} \tag{1}$$

where t = total number of time steps in the calibration period; O_i = observed value at time step i; $\bar{O}$ = average of observed value; P_i = predicted value at time step i; and $\bar{P}$ = average of predicted value.

The model performances at two different time periods (T1 and T2) were calculated in all the simulations. The acceptable performances [i.e., NSE > 0.36 (Moriasi et al. 2007)] are shown as box plots in Fig. 3(a). The mean and range of acceptable performances were similar in T1 and T2, which suggested that the overall performances were not significantly improved or reduced after selecting different time periods of rainfall event. The scatter plot shows the variability of performances in T1 and T2 [Fig. 3(b)]. There are dense dots scattered around the 1:1 line, suggesting that if it generated satisfactory simulation at T1, the set of parameters would probably generate satisfactory simulation results at T2 as well. There are also some dots far from the 1:1 line, which implies that

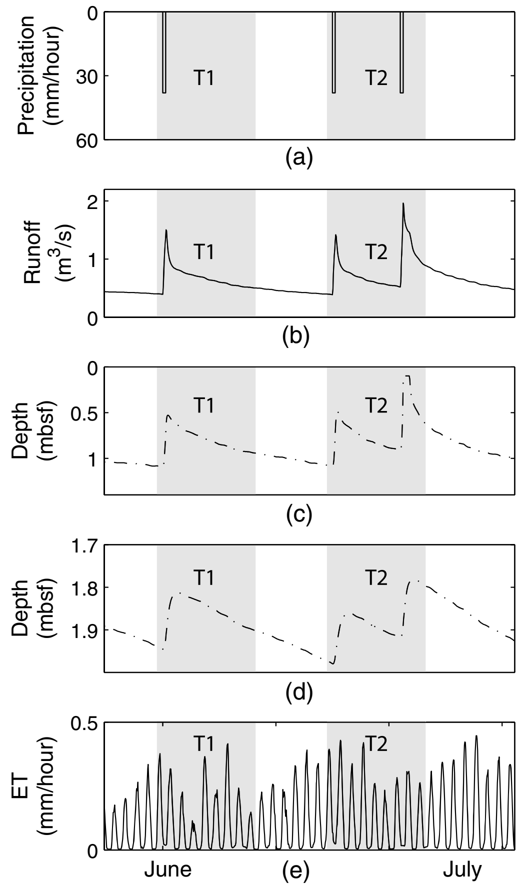

Fig. 2. Synthetic simulation results of V-shaped catchment: (a) synthetic precipitation; (b) Q; (c) GWa; (d) GWb; (e) ET (note: Q = discharge rate at the outlet; GWa = groundwater table depth at site A; GWb = groundwater table depth at site B)

there was different information contained in the simulation in T1 and T2. From the practical viewpoint, it is feasible to cut down the calibration period for computational efficiency, though some information would be lost. In the following calibration steps, only one rainfall-runoff event was simulated.

To determine the importance of each target, the model performances during T1 were evaluated. Only satisfactory performances [NSE > 0.36 (Moriasi et al. 2007)] are shown in Fig. 4(a). The well-correlated model performance implied the reciprocal relation between the two targets. The poorly correlated model performance implied competing relation. Another nine groups of Monte Carlo experiments of 10,000 PIHM simulations were performed to test the variability of the correlation. There were 10 groups of Monte Carlo simulation in total, and 10 correlations for each pair of targets [i.e., Q, GWa, GWb, and evapotranspiration (ET)], which is shown in the box plot in Fig. 4(b). Because it was stable among the 10

Table 1. Soil Hydraulic Property Characteristics for the V-Shaped Catchment

Description	Saturated hydraulic conductivity	Saturated water content	Residual water content	van Genuchten (1980) parameters: Air-entry suction α	van Genuchten (1980) parameters: Pore size distribution β	Macropore conductivity
Value	11.405	0.428	0.0339	1.31	1.72	11,405
Unit	m day^{-1}	cm^3 cm^{-3}	cm^3 cm^{-3}	m^{-1}	—	m day^{-1}

Note: Bordenan sand data are obtained from UNSODA (Leij 1996); the soil code defined in UNSODA is 4,661; the macropore conductivity is estimated as 1,000 times the saturated hydraulic conductivity.

Table 2. PIHM Parameters and Their Feasible Ranges

Parameter	Hydrological processes	Estimation	Range
Matrix conductivity	Subsurface flow	Pedotransfer functions from soil texture; field data[a]	Two orders of magnitude (multiply by 0.01–100)
Macropore conductivity	Subsurface flow	Hard coded to be 1,000 times matrix conductivity[a]	Two orders of magnitude (multiply by 0.01–100)
Topsoil conductivity	Infiltration	Pedotransfer functions from topsoil texture; field data	Two orders of magnitude (multiply by 0.01–100)
Macropore depth	Subsurface flow	Estimated from root system	0 to the bedrock depth
Porosity	Subsurface flow	Pedotransfer functions from soil texture; field data[a]	0–1
Air-entry suction α	Subsurface flow, recharge	Pedotransfer functions from soil texture	One order of magnitude (multiply by 0.1–10)
Pore size distribution β	Subsurface flow, recharge	Pedotransfer functions from soil texture	One order of magnitude (multiply by 0.1–10)
River bed conductivity	Channel routing	Hard coded to be 1.0 (vertical) and 0.1 lateral m/day[a]	Two orders of magnitude (multiply by 0.01–100)
River Manning's roughness	Channel routing	Dingman (2002)	One order of magnitude (multiply by 0.1–10)

[a]Parameters have both vertical and lateral values.

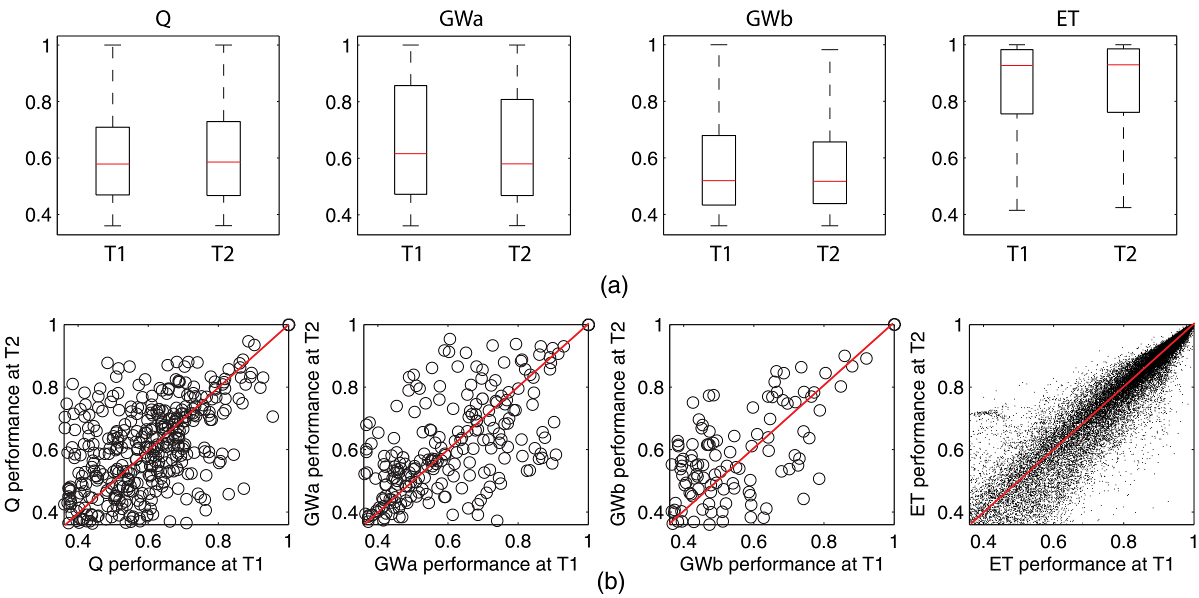

Fig. 3. Satisfactory PIHM performance during T1 and T2: (a) box plot of PIHM performance during T1 and T2; (b) scatter plot of PIHM performance during T1 and T2; the diagonal line is the 1:1 line

Monte Carlo experiments, the correlation could be used to formulate the weighted objective function. Here, informativeness was defined as the indicator of the weights of each process in the calibration. The highest level of informativeness suggested that the fitting of the particular observed variable is the priority of the calibration. The primary objective of most hydrological models is the prediction of streamflow. Hence, the value of 1 was prescribed as the informativeness of streamflow in the weighted calibration strategy. The informativeness of other variables is defined by the model performance correlation with the streamflow. Finally, the multiobjective optimization problem could be solved for a physics-based, coupled surface-subsurface, computationally expensive model by a weighted objective calibration strategy.

Weighted Function for Multiobjective Optimization

A multiobjective calibration involves the simultaneous optimization of m model residuals with respect to a vector of model parameters $\boldsymbol{\theta}$ (Gupta et al. 1998), which can be stated as

$$\min E(\boldsymbol{\theta}) = [e_i(\boldsymbol{\theta}), \ldots, e_m(\boldsymbol{\theta})], \quad \boldsymbol{\theta} \in \Theta \tag{2}$$

where the goal is to find values for $\boldsymbol{\theta}$ (a set of model parameters) within the feasible parameters space Θ that minimizes the vector $[E(\boldsymbol{\theta})]$ of multiple objectives. The multiple objectives are usually the model residuals $e_i(\boldsymbol{\theta}), i = 1, 2, 3, \ldots, m$ at each calibration objective. Here, the aggregation scheme was applied to solve the multiobjective optimization problem. The targeting variables in multiobjective functions are aggregated into a single objective with appropriate weights

$$e(\boldsymbol{\theta}) = \frac{\sum_{i=1}^{m} \omega_i \times e_i(\boldsymbol{\theta})}{\sum_{i=1}^{m} \omega_i} \tag{3}$$

where ω_i = informativeness of each calibration objective. When the information is limited to decide the appropriate weights for each objective, the unweighted case can be used with 1 as the value of informativeness. The unweighted objective function is expressed by the following equation:

$$e(\boldsymbol{\theta}) = \frac{1-\mathrm{NSE}_{\text{stream flow}} + 1-\mathrm{NSE}_{\text{site A}} + 1-\mathrm{NSE}_{\text{site B}} + 1-\mathrm{NSE}_{\mathrm{ET}}}{1+1+1+1} \tag{4}$$

The informativeness could be obtained from the previous virtual experiment (Fig. 4). Applying the informativeness values of 1, 0.9094, 0.7260, and 0.2753, the final weighted objective function turned to be

$$e(\boldsymbol{\theta}) = \frac{1-\mathrm{NSE}_{\text{stream flow}} + 0.9094\times(1-\mathrm{NSE}_{\text{site A}}) + 0.7260\times(1-\mathrm{NSE}_{\text{site B}}) + 0.2753\times(1-\mathrm{NSE}_{\mathrm{ET}})}{1+0.9094+0.7260+0.2753} \tag{5}$$

Application at Shale Hills

Catchment

The Shale Hills watershed is located in central Pennsylvania (Fig. 5). The watershed is an upland, erosion-cut, deep V-shaped valley watershed with an underlying low-permeability shale geology layer (Lynch 1976). This forested watershed supports an ephemeral stream flow, which flows into Shavers Creek, and then to the Juniata River, and ultimately the Susquehanna River. The Shale Hills watershed has been used as an experiment field in a series of hydrological studies (Nutter 1964; Lynch 1976; Qu and Duffy 2007). Recently, the watershed joined the Critical Zone Observatory (CZO) project supported by U.S. National Science Foundation as Susquehanna-Shale Hills CZO (SSHCZO).

PIHM parameterization requires the soil parameters and vegetation parameters. The soil hydraulic parameters for the van Genuchten (1980) model, including parameters to describe the inverse of air-entry suction (α), pore size distribution (β), saturated hydraulic conductivity (K_{sat}), saturated water content (h_{sat}), and residual water content (h_{res}), can be derived from the field data or estimated from soil texture (Yu et al. 2013). In this study, the soil hydraulic parameters were obtained from field data (Lin 2006). The land surface parameters such as leaf area index (LAI) and roughness length were projected from National Land Data Assimilation Systems (NLDAS 1999) vegetation parameters.

For high-resolution hydrological modeling, hourly meteorological data are required for precipitation, air temperature, relative humidity, and wind speed. These data are available from the weather station at the north ridge with 10-min frequency (Duffy 2012).

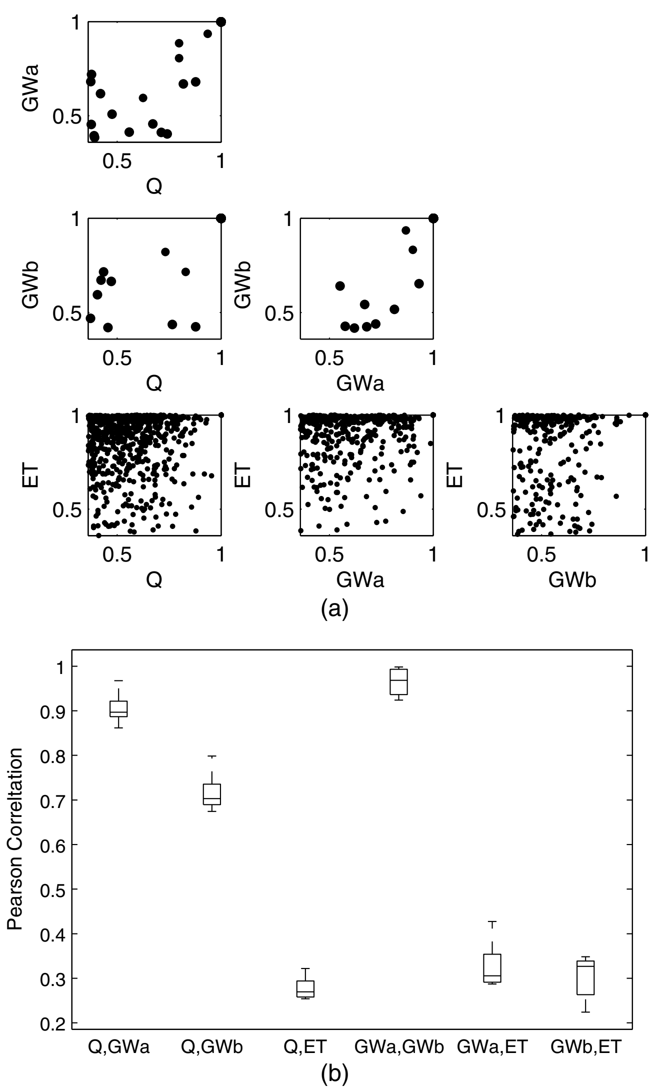

Fig. 4. NSE of PIHM performances at different variables and Pearson correlation coefficients of these NSE at 10 groups: (a) one group (10,000 simulations) of simulation results of NSE; (b) box plot of the 10 correlations from each group

Model Setup and Parameterization

PIHMgis (Bhatt et al. 2014), a tightly coupled geographic information system (GIS) interface, was applied to set up the modeling at SSHCZO. The procedures are illustrated in Fig. 6. The 1-m digital elevation model (DEM) (Guo 2010) was applied to decompose the watershed into 535 triangles and 20 linear segments of stream channels. The tree survey data (Eissenstat 2008) were used to spatially parameterize the land cover at SSHCZO. The soil classes (Lin 2006) were also projected on each computational unit of PIHM.

In the model, soil hydraulic properties are adjusted during the calibration. The estimation methods of each parameter are presented in Table 2. The range of each parameter is estimated according to the initial value. For example, the saturated hydraulic conductivity can range two orders of magnitude, while the porosity ranges from 0 to 1, and the macropore depth ranges from 0 to the bedrock depth.

There were two major flooding events in 2009. One was in June, and the other was in October. The parameters were calibrated against the flooding process in June of 2009, and the event in October was used for validation. There were observations including the streamflow at the outlet (Duffy 2010b), the water-table depths at the riparian zone (Duffy 2010a), and the upslope area (Lin 2010), and the total ET of the watershed (Davis 2010) (Fig. 5). The streamflow, water-table depths, and ET were averaged into hourly time series as the calibration objectives (Fig. 7).

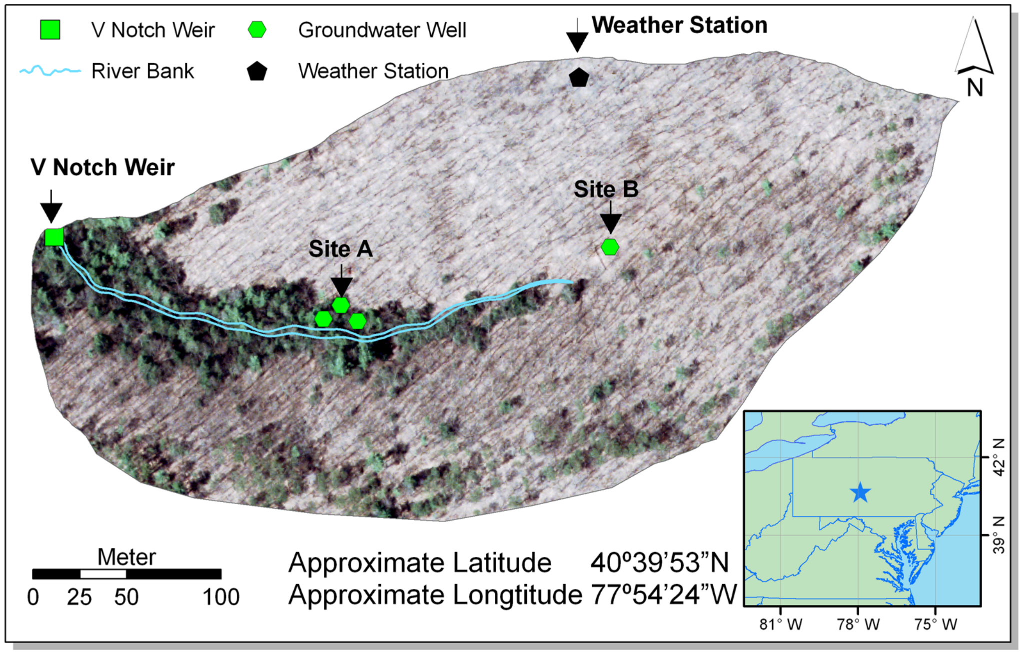

Fig. 5. Map of the SSHCZO catchment showing the locations of streamflow station, weather station, and sites of pressure transducers (map data from Shale Hills Datasets 2016)

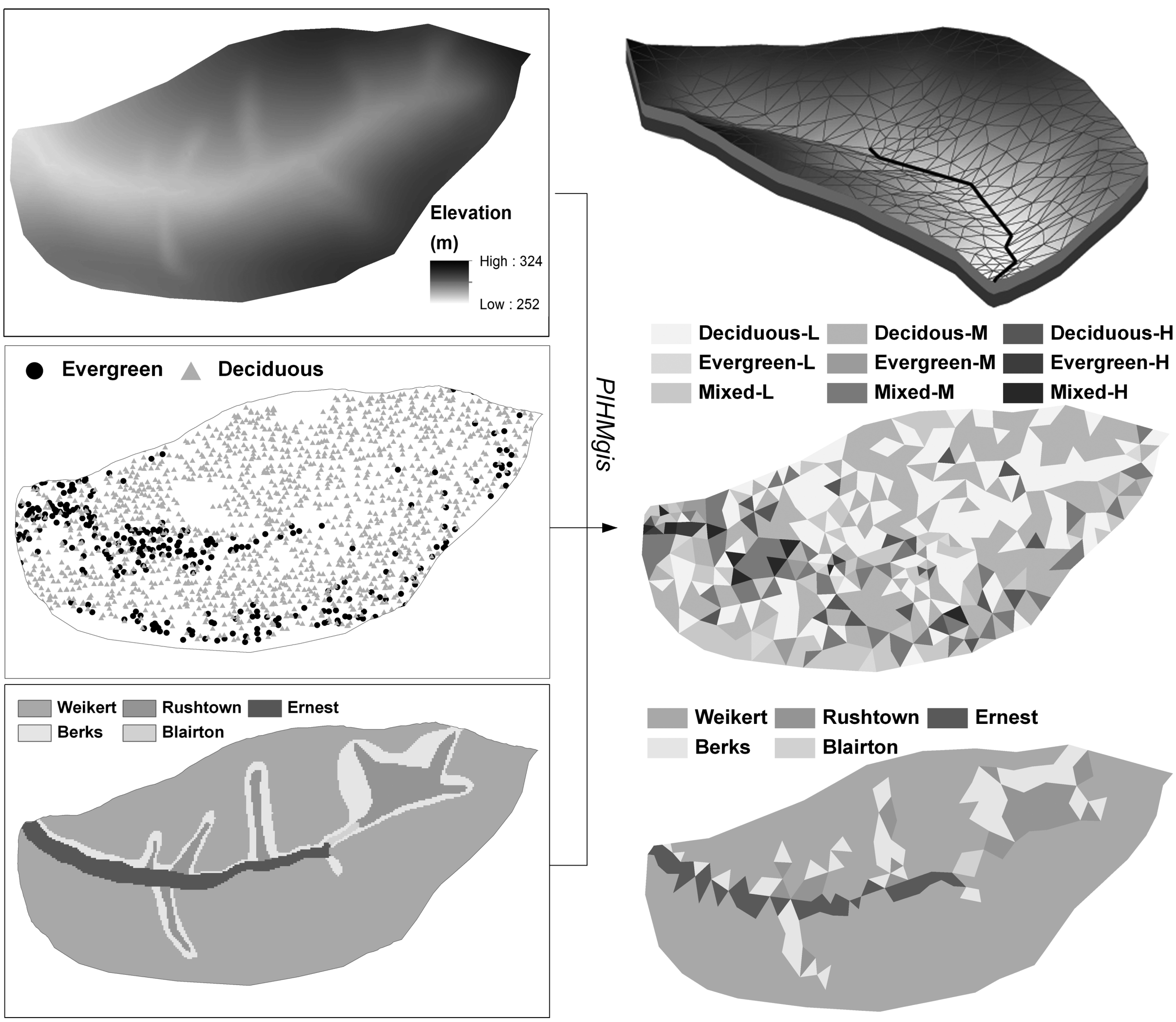

Fig. 6. PIHMgis tools process spatial layers to assign physical values and hydrological parameters

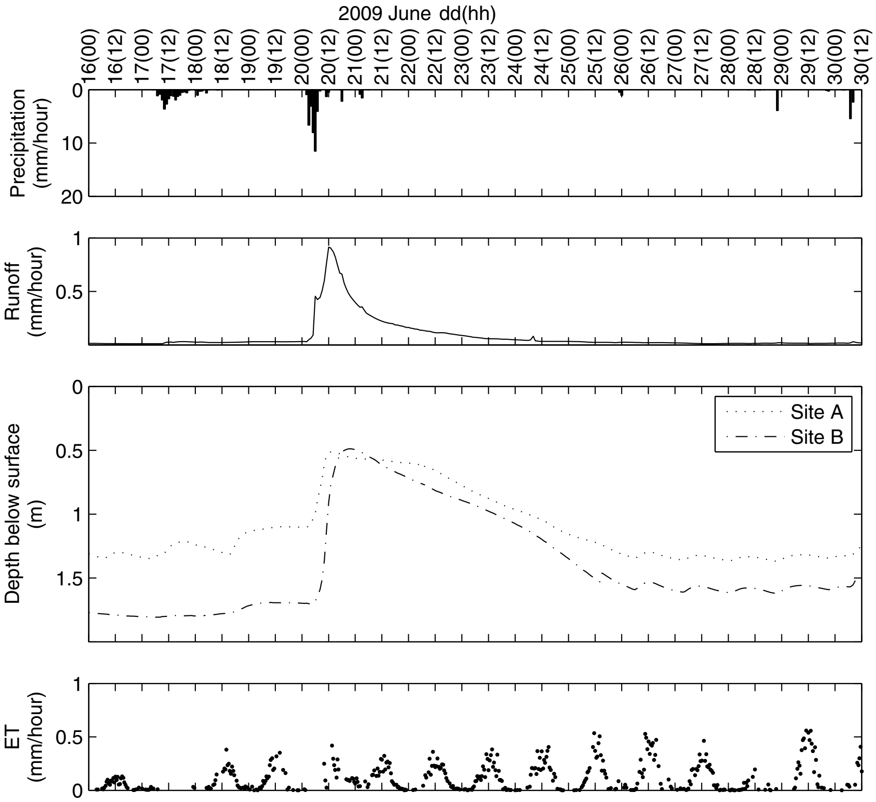

Fig. 7. Observed time series data for model calibration in June 2009; the *x*-axis is labeled in the format of date (h)

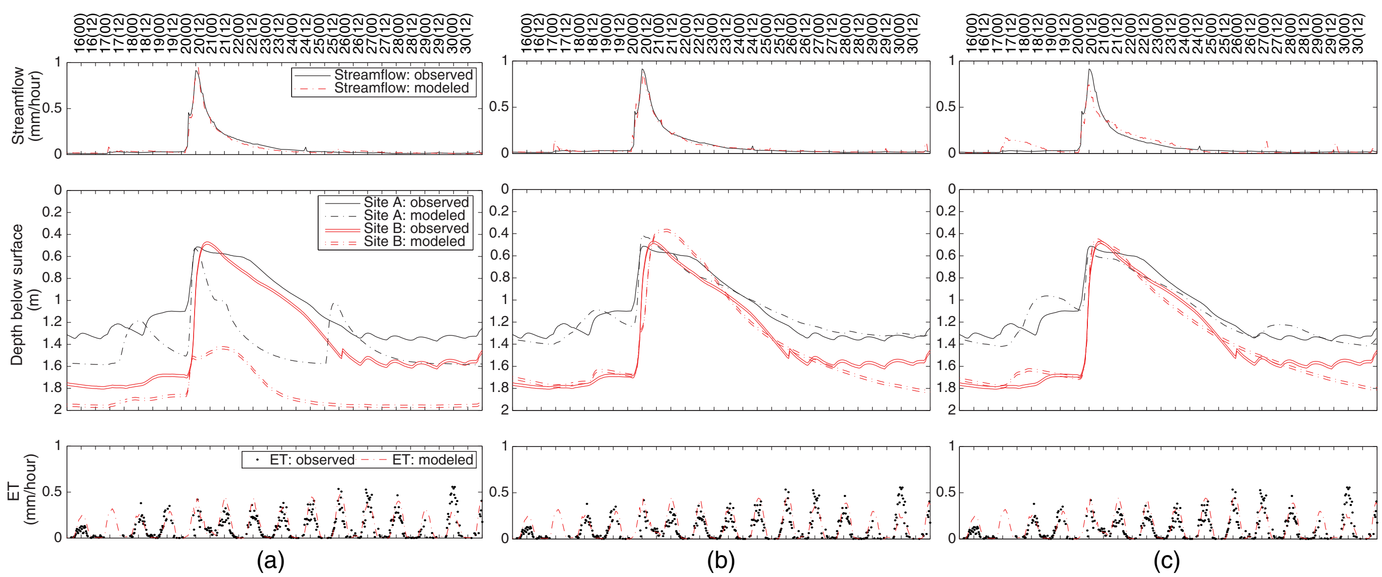

Fig. 8. Simulated and observed streamflow, groundwater table depth at Site A, Site B, and ET in the calibration period June 2009; the *x*-axis is labeled in the format of date(hour): (a) single-objective calibration; (b) informativeness-based weighted average calibration; (c) unweighted average calibration

Model Calibration and Evolution of the Model Performance

Applying the informativeness-based, weighted objective function, PIHM was calibrated against hourly streamflow, water-table depth, and ET (Fig. 8, Table 3). The covariance matrix adaptation-evolution strategy (CMA-ES) was used as the search tool (Yu et al. 2013). CMA-ES is quasi-parameter-free with the population size being the only parameter to be tuned by the user (Hansen et al. 2003). To evaluate the biases of model performance, the average of the pair differences between observed and modeled variables was calculated

Table 3. Results from Single-Objective, Informativeness-Based Weighted Average, and Unweighted Average Calibration

Objective function	Objectives	Weight in objective function	NSE	$\bar{d}$	Standard error	Z-test statistic	99% confidence interval [−2.576, 2.576]
Single-objective	Streamflow	1	0.977	5.955	2.3724	2.510	Accept
	Groundwater Site A	0	−0.867	−0.295	0.0119	−24.748	Reject
	Groundwater Site B	0	−0.927	−0.470	0.0144	−32.611	Reject
	ET	0	0.457	−1.662	0.1972	−8.430	Reject
Informativeness-	Streamflow	1	0.972	−1.363	2.5657	−0.531	Accept
based	Groundwater Site A	0.9094	0.919	−0.003	0.0041	−0.683	Accept
weighted average	Groundwater Site B	0.7260	0.866	−0.026	0.0074	−3.514	Reject
	ET	0.2753	0.460	−0.571	0.247	−2.313	Accept
Unweighted average	Streamflow	1	0.880	−11.793	5.3561	−2.202	Accept
	Groundwater Site A	1	0.877	−0.017	0.0058	−2.897	Reject
	Groundwater Site B	1	0.914	−0.012	0.0049	−2.490	Accept
	ET	1	0.593	−1.646	0.1947	−8.451	Reject

Note: The simulation period is June 2009; the time series are shown in Fig. 8.

$$
\begin{array}{ll}
N: & \text{sample size} \\
\bar{d}: & d_i = O_i - P_i \\
\text{SD}: & \sqrt{\frac{\sum_{i=1}^{N}(d_i-\bar{d})^2}{N}} \\
\text{SE}: & \frac{\text{SD}}{\sqrt{N}} \\
Z\text{-test}: & \frac{\bar{d}-0}{\text{SE}}
\end{array}
\tag{6}
$$

where N = sample size; O_i = observed variable at time i; and P_i = model simulated variable at time i; SD = standard deviation; and SE = standard error. The value of d_i is calculated as the sample difference between O_i and P_i. The hypotheses was set as H_0: $= 0$ (i.e., the difference between the obeserved and modeled variable equals zero), H_1: $\neq 0$ (i.e., the difference between the observed and modeled variable does not equal 0). And then the Z-test was used to test the hypothses. Here, 0.01 was used as the significance level. Therefore, the 99% confidence interval was [−2.576, 2.576]. If the Z-test statistic belonges to the confidence interval, the hypothesis is acceptable: the obeserved and modeled variables are not signicantly different from each other.

The calibration by the weighted-objective function improved the prediction of water-table depths and ET without significant degredation of the streamflow prediction. According to the Z-test results (Table 3), both streamflow and groundwater table at Site A and ET obtained good model performance, i.e., the modeled values were not siginificantly different from the observed values. Without informativeness, the unweighted averaged calibration was inclined to the prediction of the groundwater table at Site B and ET (Fig. 8). The Z-test results suggested that only the streamflow and groundwater table at Site B obtained acceptable model performance (Table 3).

The parameters from the weighted objective calibration performed continously well in the validation periods. Three sets of parameters [from single objective (with Q only), informativeness-based weighted average, and unweighted average] were applied to simulate the flood event in October 2009 (Fig. 9, Table 4). The single-objective case had the best performace at streamflow, while the other two cases improved the performances at groundwater table and ET. The unweighted calibration overemphasized the streamflow prediction. The informativeness-based weighted average case demonstrated a reasonable compromise between the objectives. The Z-test results suggested that both streamflow and groundwater table at Site A persisted acceptable model performance (i.e., observed and modeled variables are not signicantly different from each other), though the NSE values of the groundwater table at Site A suggeted poor performance. The validation results of ET were not as good as other objectives in all the simulations. The vegetation parameters were not calibrated in this study and the ET was calculated from the eddy covariance measurements by a Campbell Scientific CSAT3 three-dimensional sonic anemometer (3D Sonic Anemometer, Campbell Scientific, Logan, Utah) and a LI-COR LI-7500 CO_2/H_2O analyzer (LI-COR, Lincoln, Nebraska). The method itself includes unavoidable measurement uncertainty.

Discussion

Significance of Benchmarking Simulation

Bencharking simulations are easily reproducible and reusable to understand model parameters. During the development of PIHM, the V-shaped catchment was only used for the validation of PIHM (Qu 2005). After that, real-world watershed developments and applications of PIHM rarely paid attention to the V-shaped catchment case. The authors argue that simplified catchment processes lead to generalizable insights and understanding. The V-shaped catchment was extended with subsurface domain and meteorological input to understand the calibration processes of PIHM. Such documentation of the V-shaped catchment can be easily repeated, reproduced, and reused by other modelers and other coupled surface-subsurface models. Maxwell et al. (2014) listed a series of benchmarking cases for coupled surface-subsurface models to understand the differences of each model. Another simple experiment is the Henry problem (Henry 1964), which has been studied repeatedly to understand coastal groundwater intrusion. The authors' virtual experiment incorporated subsurface and evapotranspiration, which has potential of reuse for understanding the integrated watershed hydrologic processes. This work represents a first step in decomposing the calibration targets, which may improve the ability to utilize simple generalization to understand complex system (Wainwright and Mulligan 2004).

Use Event-Scale Data for Calibration

The calibration period depends on the structure of the model and characteristic of each parameter. Traditionally, the model calibration period requires a longer time period than validation because some parameters do not have phyiscal meanings or cannot be extracted from physical properties of the model domain. The reason that event-scale hourly data were enough for this study is the physics-based characteristic of the parameters, the spatial scale of Shale Hills watershed, and the hydrologic dynamics of PIHM simulation.

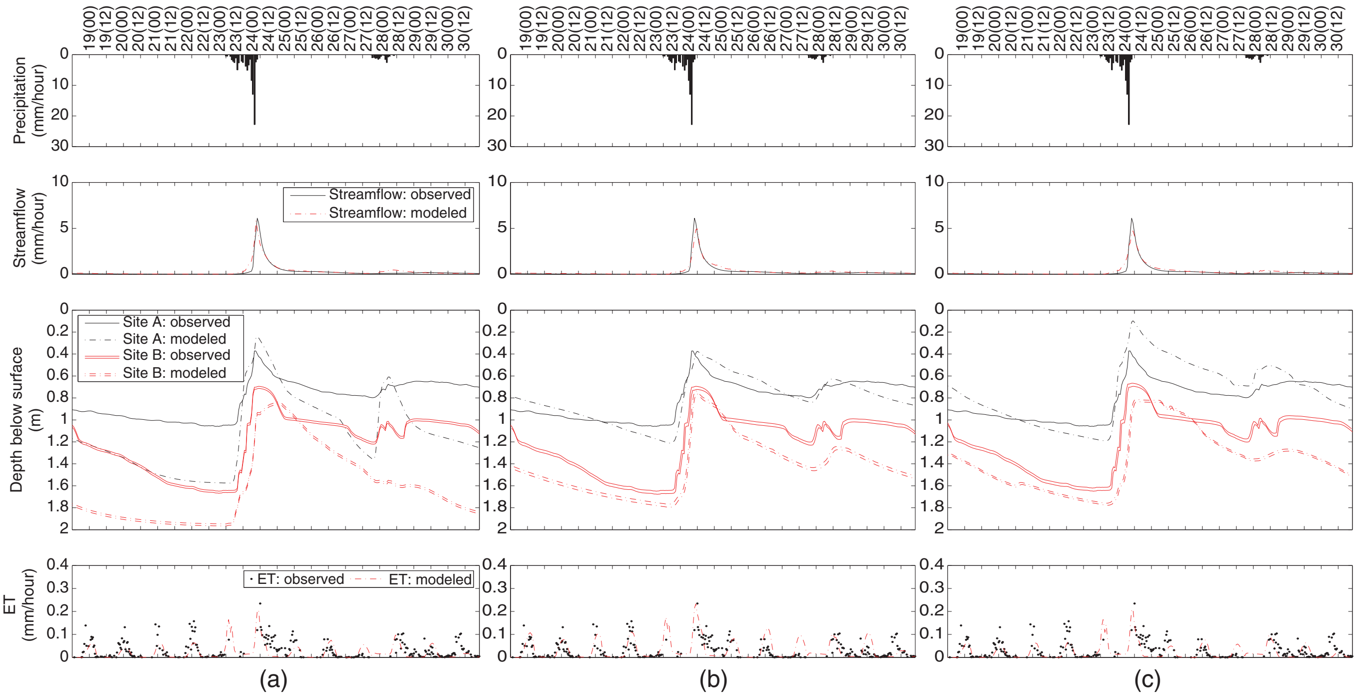

Fig. 9. Simulated and observed streamflow, groundwater table depth at Site A, Site B, and ET in the validation period October 2009; the *x*-axis is labeled in the format of date (h): (a) single-objective calibration; (b) informativeness-based weighted average calibration; (c) unweighted average calibration

Table 4. Validation Results from Single-Objective, Informativeness-Based Weighted Average, and Unweighted Average Calibration

Objective function	Objectives	Weight in objective function	NSE	$\bar{d}$	Standard error	*Z*-test statistic	99% confidence interval [−2.576, 2.576]
Single-objective	Streamflow	1	0.936	−38.715	25.160	−1.539	Accept
	Groundwater Site A	0	−5.108	−0.173	0.012	−14.139	Reject
	Groundwater Site B	0	−2.940	−0.223	0.013	−17.828	Reject
	ET	0	−0.856	1.081	0.181	5.981	Reject
Informativeness-based weighted average	Streamflow	1	0.927	−45.180	22.370	−2.020	Accept
	Groundwater Site A	0.9094	−0.069	−0.009	0.006	−1.603	Accept
	Groundwater Site B	0.7260	0.056	−0.127	0.007	−18.200	Reject
	ET	0.2753	−0.856	0.954	0.189	5.041	Reject
Unweighted average	Streamflow	1	0.816	−46.195	20.780	−2.223	Accept
	Groundwater Site A	1	0.5956	0.044	0.009	4.753	Reject
	Groundwater Site B	1	0.826	−0.098	0.008	−11.747	Reject
	ET	1	−0.850	1.073	0.181	5.919	Reject

Note: The simulation period is October 2009; the time series are shown in Fig. 9.

One rainfall-runoff response event includes a complete set of hydrological processes including infiltration, surface runoff, subsurface flow, and stream channel flow. During the event-scale calibation, PIHM parameters capture the phyisical meaning, which drives the coupling dynamics of these processes. These dynamics are memorized in the parameters so that the model can still be effective to predict other rainfall-runoff responses.

Meaning of Informativeness of Each Calibration Objective

Informativeness represents transferable information on parameters between different watersheds. The idea of calibration is to transfer model parameters temporally. Here, the real-world watershed calibration represents transferable information from a simple watershed. This study demonstrated that it is effective when the transfer happens on similar catchment structure and similar target variables. Future studies should explore other indirect transfering methods between real-world watersheds.

The implementation of integrated distributed environmental models certainly increases the amount of observational data required to constrain the model parameters (Stisen et al. 2011). For PIHM, the calibration strategy suggests that streamflow is strongly dependent on the water-table depth close to stream (GWa, Fig. 1). In Fig. 4(a), it can be seen that the model performances at GWa and Q were well correlated with each other: i.e., changing a parameter set will either increase or decrease the NSE of GWa and Q. However, the model performance of GWb was weakly correlated with the performance of streamflow. The low correlation of model performance between ET and other variables suggests that the model performance for an event does not significantly affect the ET process simulation (Yu et al. 2013; Shi et al. 2013). In addtion, the observation uncertainty may affect the informativeness, and low informativenss of ET may suggest the large uncertainty of ET observation, which should be further scrutinized.

The informativeness concept reflects a model coupling scheme of different processes and enables consideration of the multiple constraints of the watershed. The unweighted method showed that the calibration mistakenly focused on the ET and the water table at Site B, and failed to reproduce the observed outlet streamflow. The informativeness-based strategy of formulating the weighted objective function avoids subjective judgements and could be easily adopted by other integrated models. This study used streamflow as the dominant objective. The informativeness will be different if another dominant objective is applied accroding to the prescribed simulation priority.

Strength of the Informativeness-Based Weighted Calibration

With the weighted constraint of multiple measurements, the PIHM simulation results significantly enhanced the prediction of water-table depths. The key aims of distributed modeling schemes are to reproduce multiple moisture fluxes and to reflect the spatial heterogeneities of the hydrological mechanisms (Kim et al. 2012). Here, the authors improved the representation of multiple processes by weighted average calibration, and the weighted average function can be easily extended to more constraints of the model as the types of observed variables increase.

Calibration with informativeness-based weights directly generate one optimum solution. Multiobjective optimization problems can be resolved by nondominated sorting. This method often generates a large number of Pareto optimal sets. Ususally, a further step of selection is necessary to obtain a small number of Pareto fronts as physically sound solutions (Khu and Madsen 2005). It might be a challenge to select a final set of parameters from the nondomainted solutions. In this study, the informativeness-based weights decide the final selection of the solution. The authors argue that the informativeness represents a preference for ordering each calibration objective in the integrated hydrologic model. The weighted average aggregation method reduced the computational cost and avoided the selection process from a large number of Pareto-optimal solutions, which is efficient for computationally expensive models.

Conclusion

This paper develops a benchmarking simulation of PIHM on the V-shaped catchment. The simulation on V-shaped catchment implies the potential of using event-scale period data to calibrate PIHM. In addition, the model performance correlations of different targets are proved to be stable in 10 groups of Monte Carlo simulations. It implies that model performance correlations can be used as the aggregation for a weighted objective function. Therefore, the authors tested the informativeness-based, weighted objective calibration of PIHM using observed streamflow, water-table depth, and ET. The following conclusions are drawn from this paper:

- The virtual experiment on a V-shaped catchment can efficiently guide the calibration of PIHM by formulating the weighted objective function according to the model performance correlations.
- Selection of different rainfall-runoff events does not improve the PIHM simulation results. Therefore, a simple rainfall-runoff event can be used for PIHM calibration.
- The informativeness provided a useful framework for objectively determining weights between each calibration objective. Here, the authors prescribed streamflow as the dominant objective, and the correlations of model performance at other objectives were used for the evaluation of informativeness. Results suggest a satisfactory compromise among streamflow, water-table depth, and ET was achieved with the weighted strategy.
- The comparison between single-objective optimization and weighted multiobjective optimization suggests that sound coupled surface-subsurface modeling relies on appropriate calibration against multiple observations.

The proposed calibration framework may be categorized under the umbrella of physics-based models with comprehensive parameter representation of topography, land cover, soil, and geology. Most of the physics-based parameters would not change significantly in different hydrologic conditions, and hold relatively long stability. Therefore, it is practical to shorten the calibration duration by picking out less important information. Here, the direct transfer of the model performance correlations is feasible due to a similar domain between the V-shaped catchment and the Shale Hills watershed. Spatially, the authors hope that these physics-based parameters have more transferable information. It would be valuable to explore other strategies to transfer information across neighboring or similar catchments with diverse complexity and scales.

Acknowledgments

The authors thank the editor and reviewers for their valuable input, which helped to improve the manuscript. This research was supported by the grant from the National Science Foundation, EAR-0725019, EAR-1239285, and EAR-1331726 Shale Hills-Susquehanna Critical Zone Observatory. The computing facility was supported in part through instrumentation funded by the National Science Foundation through Grant OCI–0821527.

References

Bhatt, G., Kumar, M., and Duffy, C. J. (2014). “A tightly coupled GIS and distributed hydrologic modeling framework.” *Environ. Modell. Software*, 62, 70–84.

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

Chen, F., and Dudhia, J. (2001). “Coupling an advanced land surface-hydrology model with the Penn State-NCAR MM5 modeling system. Part I: Model implementation and sensitivity.” *Mon. Weather Rev.*, 129(4), 569–585.

Davis, K. (2010). “CZO dataset: Shale Hills—Flux tower (2009–2010).” ⟨http://criticalzone.org/shale-hills/data/dataset/2558/⟩ (Feb. 23, 2012).

Di Giammarco, P., Todini, E., and Lamberti, P. (1996). “A conservative finite elements approach to overland flow: The control volume finite element formulation.” *J. Hydrol.*, 175(1–4), 267–291.

Dingman, S. L. (2002). *Physical hydrology*, 2nd Ed., Prentice Hall, Upper Saddle River, NJ.

Duffy, C. J. (2010a). “CZO dataset: Shale Hills—Groundwater depth data (2009–2012).” ⟨http://cataract.cee.psu.edu/czo/rth2/index.php?dir=Groundwater⟩ (Feb. 24, 2012).

Duffy, C. J. (2010b). “CZO dataset: Shale Hills—Streamflow data (2006–2012).” ⟨http://cataract.cee.psu.edu/czo/rth2/index.php?dir=Streamflow⟩ (Feb. 24, 2012).

Duffy, C. J. (2012). “CZO dataset: Shale Hills—Precipitation (2006–2012).” ⟨http://criticalzone.org/shale-hills/data/dataset/2556/⟩ (Feb. 24, 2012).

Dung, N. V., Merz, B., Bárdossy, A., Thang, T. D., and Apel, H. (2011). “Multi-objective automatic calibration of hydrodynamic models utilizing inundation maps and gauge data.” *Hydrol. Earth Syst. Sci.*, 15(4), 1339–1354.

Efstratiadis, A., and Koutsoyiannis, D. (2010). "One decade of multi-objective calibration approaches in hydrological modelling: A review." *Hydrol. Sci. J.*, 55(1), 58–78.

Eissenstat, D. (2008). "CZO dataset: Shale Hills—Vegetation—Tree survey." ⟨http://criticalzone.org/shale-hills/data/dataset/2648/⟩ (Feb. 24, 2012).

Guo, Q. (2010). "Susquehanna Shale Hills critical zone observatory: Leaf off survey." San Diego Supercomputer Center, University of California San Diego, La Jolla, CA.

Gupta, H. V., Sorooshian, S., and Yapo, P. O. (1998). "Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information." *Water Resour. Res.*, 34(4), 751–763.

Hansen, N., Muller, S. D., and Koumoutsakos, P. (2003). "Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES)." *Evol. Comput.*, 11(1), 1–18.

Henry, H. R. (1964). "Effects of dispersion on salt encroachment in coastal aquifers." U.S. Geological Survey, 70–84.

Hsie, M., Yan, S., and Pan, N. (2014). "Improvement of rainfall-runoff simulations using the runoff-scale weighting method." *J. Hydrol. Eng.*, 10.1061/(ASCE)HE.1943-5584.0000921, 1330–1339.

Juston, J., Seibert, J., and Johansson, P.-O. (2009). "Temporal sampling strategies and uncertainty in calibrating a conceptual hydrological model for a small boreal catchment." *Hydrol. Processes*, 23(21), 3093–3109.

Khu, S. T., and Madsen, H. (2005). "Multiobjective calibration with Pareto preference ordering: An application to rainfall-runoff model calibration." *Water Resour. Res.*, 41(3), W03004.

Kim, J., Warnock, A., Ivanov, V. Y., and Katopodes, N. D. (2012). "Coupled modeling of hydrologic and hydrodynamic processes including overland and channel flow." *Adv. Water Resour.*, 37(0), 104–126.

Kumar, M. (2009). "Toward a hydrologic modeling system." Ph.D. dissertation, Pennsylvania State Univ., University Park, PA.

Kumar, M., and Duffy, C. (2015). "Exploring the role of domain partitioning on efficiency of parallel distributed hydrologic model simulations." *J. Hydrogeol. Hydrol. Eng.*, 4(1), 1–12.

Leij, F. J. (1996). "The UNSODA unsaturated soil hydraulic database: User's manual." U.S. Environmental Protection Agency, Ada, Oklahoma.

Li, X., Weller, D. E., and Jordan, T. E. (2010). "Watershed model calibration using multi-objective optimization and multi-site averaging." *J. Hydrol.*, 380(3), 277–288.

Lin, H. (2006). "Temporal stability of soil moisture spatial pattern and subsurface preferential flow pathways in the Shale Hills catchment." *Vadose Zone J.*, 5(1), 317–340.

Lin, H. (2010). "CZO dataset: Shale Hills—Hydropedologic properties (2007–2010)—Water table." ⟨http://criticalzone.org/shale-hills/data/dataset/2586/⟩ (Feb. 23, 2012).

Lynch, J. A. (1976). "Effects of antecedent soil moisture on storm hydrographs." Ph.D. thesis, Pennsylvania State Univ., University Park, PA.

Mallard, J., McGlynn, B., and Covino, T. (2014). "Lateral inflows, stream groundwater exchange, and network geometry influence stream water composition." *Water Resour. Res.*, 50(6), 4603–4623.

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(2), 1531–1549.

McMillan, H., Krueger, T., and Freer, J. (2012). "Benchmarking observational uncertainties for hydrology: Rainfall, river discharge and water quality." *Hydrol. Processes*, 26(26), 4078–4111.

Miller, N. L. (1995). "Sensitivity of surface heat and moisture fluxes due to topographic slope and azimuth." *J. Geophys. Res. Atmos.*, 100(D9), 18669–18685.

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., and Veith, T. L. (2007). "Model evaluation guidelines for systematic quantification of accuracy in watershed simulations." *Trans. ASABE*, 50(3), 885–900.

Nicklow, J., et al. (2010). "State of the art for genetic algorithms and beyond in water resources planning and management." *J. Water Resour. Plann. Manage.*, 10.1061/(ASCE)WR.1943-5452.0000053, 412–432.

NLDAS (National Land Data Assimilation Systems). (1999). "Mapped monthly vegetation parameters." ⟨http://ldas.gsfc.nasa.gov/nldas/NLDASmapveg.php⟩ (Mar. 10, 2012).

Nutter, W. L. (1964). "Determination of the head-discharge relationship for a sharp-crested compound weir and a sharp-crested parabolic weir." Ph.D. thesis, Pennsylvania State Univ., University Park, PA.

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

Persson, P.-O., and Strang, G. (2004). "A simple mesh generator in MATLAB." *SIAM Rev.*, 46(2), 329–345.

PIHM (Penn State Integrated Hydrologic Modeling System). (2016). "Overview." ⟨http://www.pihm.psu.edu/⟩ (Jun. 15, 2016).

Qu, Y. (2005). "An integrated hydrologic model for multi-process simulation using semi-discrete finite volume approach." Pennsylvania State Univ., University Park, PA.

Qu, Y., and Duffy, C. J. (2007). "A semidiscrete finite volume formulation for multiprocess watershed simulation." *Water Resour. Res.*, 43(8), W08419.

Razavi, S., and Tolson, B. A. (2013). "An efficient framework for hydrologic model calibration on long data periods." *Water Resour. Res.*, 49(12), 8418–8431.

Rientjes, T. H. M., Muthuwatta, L. P., Bos, M. G., Booij, M. J., and Bhatti, H. A. (2013). "Multi-variable calibration of a semi-distributed hydrological model using streamflow data and satellite-based evapotranspiration." *J. Hydrol.*, 505(15), 276–290.

Rozos, E., Efstratiadis, A., Nalbantis, I., and Koutsoyiannis, D. (2004). "Calibration of a semi-distributed model for conjunctive simulation of surface and groundwater flows." *Hydrol. Sci. J.*, 49(5), 819–842.

Seo, Y., and Schmidt, A. R. (2013). "Network configuration and hydrograph sensitivity to storm kinematics." *Water Resour. Res.*, 49(4), 1812–1827.

Shale Hills Datasets. (2012). ⟨http://criticalzone.org/shale-hills/data/datasets/⟩ (Feb. 24, 2012).

Shi, Y., Davis, K. J., Duffy, C. J., and Yu, X. (2013). "Development of a coupled land surface hydrologic model and evaluation at a critical zone observatory." *J. Hydrometeorol.*, 14(5), 1401–1420.

Shi, Y., Davis, K. J., Zhang, F., Duffy, C. J., and Yu, X. (2014). "Parameter estimation of a physically based land surface hydrologic model using the ensemble Kalman filter: A synthetic experiment." *Water Resour. Res.*, 50(1), 706–724.

Simpson, M. J., and Clement, T. P. (2003). "Theoretical analysis of the worthiness of the Henry and Elder problems as benchmarks of density-dependent groundwater flow models." *Adv. Water Resour.*, 26, 17–31.

Singh, S. K., and Bárdossy, A. (2012). "Calibration of hydrological models on hydrologically unusual events." *Adv. Water Resour.*, 38, 81–91.

Stisen, S., McCabe, M. F., Refsgaard, J. C., Lerer, S., and Butts, M. B. (2011). "Model parameter analysis using remotely sensed pattern information in a multi-constraint framework." *J. Hydrol.*, 409(1), 337–349.

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.

Vrugt, J. A., Gupta, H. V., Dekker, S. C., Sorooshian, S., Wagener, T., and Bouten, W. (2006). "Application of stochastic parameter optimization to the Sacramento soil moisture accounting model." *J. Hydrol.*, 325(1–4), 288–307.

Wainwright, J., and Mulligan, M. (2004). *Environmental modelling: Finding simplicity in complexity*, Wiley, West Sussex, U.K.

Weiler, M., and McDonnell, J. (2004). "Virtual experiments: A new approach for improving process conceptualization in hillslope hydrology." *J. Hydrol.*, 285(1–4), 3–18.

Yu, X. (2015). "PIHM input to benchmarking V-catchment." ⟨http://dx.doi.org/10.6084/m9.figshare.1328521⟩ (Apr. 11, 2015).

Yu, X., Bhatt, G., Duffy, C., and Shi, Y. (2013). "Parameterization for distributed watershed modeling using national data and evolutionary algorithm." *Comput. Geosci.*, 58, 80–90.

Zidane, A., Younes, A., Huggenberger, P., and Zechner, E. (2012). "The Henry semianalytical solution for saltwater intrusion with reduced dispersion." *Water Resour. Res.*, 48, W06533.