



RESEARCH ARTICLE

10.1002/2015EA000155

Special Section:

Geoscience Papers of the Future

Key Points:

- We document a computational workflow for implementing a coupled surface-subsurface hydrologic model
- Reproducing modeling results is achieved for a benchmark example problem and a real-world application
- Future publications should make data and software persistent, linked, user-friendly, and sustainable

Supporting Information:

- Supporting Information S1
- Data Set S1

Correspondence to:

X. Yu,
xuan@udel.edu

Citation:

Yu, X., C. J. Duffy, A. N. Rousseau, G. Bhatt, Á. Pardo Álvarez, and D. Charron (2016), Open science in practice: Learning integrated modeling of coupled surface-subsurface flow processes from scratch, *Earth and Space Science*, 3, 190–206, doi:10.1002/2015EA000155.

Received 1 FEB 2016

Accepted 17 APR 2016

Accepted article online 3 MAY 2016

Published online 9 MAY 2016

©2016. The Authors.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

Open science in practice: Learning integrated modeling of coupled surface-subsurface flow processes from scratch

Xuan Yu¹, Christopher J. Duffy², Alain N. Rousseau³, Gopal Bhatt², Álvaro Pardo Álvarez³, and Dominique Charron³

¹Department of Geological Sciences, University of Delaware, Newark, Delaware, USA, ²Department of Civil and Environmental Engineering, Pennsylvania State University, University Park, Pennsylvania, USA, ³INRS-ETE, Institut National de la Recherche Scientifique (Centre Eau Terre Environnement), Université du Québec, Québec City, Québec, Canada

Abstract Integrated modeling of coupled surface-subsurface flow and ensuing role in diverse Earth system processes is of current research interest to characterize nonlinear rainfall-runoff response and also to understand land surface energy balances, biogeochemical processes, geomorphological dynamics, etc. A growing number of complex models have been developed for water-related research, and many of these are made available to the Earth science community. However, relatively few resources have been made accessible to the potentially large group of Earth science and engineering users. New users have to invest an extraordinary effort to study the models. To provide a stimulating experience focusing on the learning curve of integrated modeling of coupled surface-subsurface flow, we describe use cases of an open source model, the Penn State Integrated Hydrologic Model, PIHM. New users were guided through data processing and model application by reproducing a numerical benchmark problem and a real-world watershed simulation. Specifically, we document the PIHM application and its computational workflow to enable intuitive understanding of coupled surface-subsurface flow processes. In addition, we describe the user experience as important evidence of the significance of reusability. The interaction shows that documentation of data, software, and computational workflow in research papers is a promising method to foster open scientific collaboration and reuse. This study demonstrates how open science practice in research papers would promote the utility of open source software. Addressing such open science practice in publications would promote the utility of journal papers. Further, popularization of such practice will require coordination among research communities, funding agencies, and journals.

1. Introduction

Hydrologic models conceptualize and simplify representations of physical processes governing the movement of water on, above, and below the Earth surface. Recent years have witnessed the emergence of complex coupled surface-subsurface models that explicitly account for the spatial physical patterns describing the continuous water fluxes between the land surface, soil or vadose zone, and the underlying aquifers (i.e., phreatic zone). The physical processes include overland flow, land cover vegetation water use, unsaturated flow, and groundwater flow. Such models (see recent literature review by *Fatichi et al.* [2016] and *Maxwell et al.* [2014]) are essential tools to understand watershed hydrology [*De Schepper et al.*, 2015], coastal salinization [*Yang et al.*, 2013], subsurface storage [*Niu et al.*, 2014], and climate change impacts assessment [*Maxwell and Kollet*, 2008], to name a few. Furthermore, due to the comprehensive and physical representation of hydrological processes, coupled surface-subsurface models are of interest to many other research communities (e.g., wetland ecosystem [*Fossey et al.*, 2015; *Golden et al.*, 2014; *Yu et al.*, 2015a] and geomorphology [*Paola et al.*, 2006; *Zhang et al.*, 2016]). Collaboration between hydrologic modelers and other scientists has proven to be a key method to answer synthetic and integrated research questions on water-related research with physical, ecological, and social sciences [*Brooks et al.*, 2015; *Clark et al.*, 2015; *Penn et al.*, 2016; *Sutanudjaja et al.*, 2015].

Coupled surface-subsurface hydrologic models provide many opportunities to engage general public and support geoscience education. Modeling of continuous and simultaneous hydrological processes requires spatially and temporally distributed measurements for validation. *Lowry and Fienen* [2013] presented that the untrained observers could collect accurate stream stage data with low cost, which implied

promise for stimulating public engagement in the hydrologic modeling studies. *Buytaert et al.* [2014] noted that the communication of different observation methods and modeling results is a major challenge in the quality control of citizen science (i.e., the participation of general public in the research design and data collection and interpretation process)-generated data. For educational purposes, many complex geoscientific concepts and phenomena can be demonstrated by the simulation results of coupled surface-subsurface hydrologic models. *Dickerson and Dawkins* [2004] showed that an appropriate groundwater model could assist students in better understanding of the scientific terms and conceptions of hydrologic cycle. *Shepardson et al.* [2007] argued that geoscience education with a conceptual watershed model could promote a citizenry that is knowledgeable about the watershed concept. These studies implied that effective communication could bring a large group of potential users of coupled surface-subsurface models. Such broad interest in these models led to the question: how do readers understand, learn, and make use of the simulation results of coupled surface-subsurface flow models presented in the publications?

Currently, the utility of coupled surface-subsurface hydrologic models is degraded by the limited access of reusable data sets and transparent simulation procedures. The procedures for generating and processing modeling results are often not transparent, which leads to barriers to the communication of knowledge and adoption of new techniques. In many hydrologic modeling studies, authors spend more time on data preparation and data visualization pipelines than running the model itself. Typically, the implementation details and simulation results for these complex models are only summarized and not transparent in the final publication, though tutorials may be available at the model development group. It is often the case that readers and potential users may not have enough information to be certain about the data sources and simulation results, let alone replicate or reuse the model simulation from the published text, figures, and tables. Therefore, data and simulation results only serve a particular research project and are hidden to scientists and other potential users. *Heidorn* [2008] used the term “dark data” to refer to the scientific data with poor utility due to ineffective index and inappropriate storage. Due to complex data variables and analysis pipelines, the fidelity of model simulation may not be convincingly conveyed through plain text [*Stodden*, 2011]. Like the famous Harlow Shapley quotation: “No one trusts a model except the man who wrote it; everyone trusts an observation, except the man who made it” [*Roulstone and Norbury*, 2013], many model simulation processes are often not transparently presented.

The publication of research papers on coupled surface-subsurface modeling studies is slowly changing, though cyberinfrastructure has been already implemented in many research facilities [e.g., *Tarboton et al.*, 2014; *Yu et al.*, 2015b]. Such projects aimed to explore community sharing on data and model and interoperability on environmental information and services, including Critical Zone Observatory (<http://criticalzone.org/>), DRIHM (<http://www.drihm.eu>), EVO (<http://www.evo-uk.org>), GEOSS (<http://www.earthobservations.org/>), HydroTerre (www.hydroterre.psu.edu), HydroShare (<http://hydroshare.cuahsi.org>), HyperHydro (<http://www.hyperhydro.org>), NEON (<http://www.neoninc.org>), and the Organic Data Science Framework (<http://www.organicdatascience.org>). Meanwhile, the research publication is still focusing on the hypotheses as the tradition. Recently, data papers [*Mirus et al.*, 2011] and data journals [*Mayernik et al.*, 2015; *Kattge et al.*, 2014] in the field of hydrology have attempted to address the importance of data documentation in modeling studies. However, documentation of the process of model implementation (e.g., source code, model versions, and platform) has not yet become a standard practice in the literature [*Skaggs et al.*, 2015].

Another overlooked issue for improving transparency and utility is to provide a record of origin (i.e., provenance) of the model simulation results, which includes all intermediate data and associated analysis pipelines. Reproducibility implies the ability to recreate the results or retest the hypotheses leading to the same conclusion obtained, whether by rerunning the same code, or by writing a new program, or by designing a new computational experiment [*Easterbrook*, 2014]. Providing complete data sets and provenance will help readers to test if they can reproduce each step of the computation, to improve the model, or to develop new methods. *Ceola et al.* [2015] reported a comparative experiment of hydrologic simulation over 15 European catchments to explore the reproducibility of computational methods in hydrology. The results revealed that the reproducibility between each group could be ensured by protocols not only on modeling purpose and input data but also on execution steps and result analysis. Without demonstration of computational data-model workflows, the readers and/or users will find it hard to use the data and software to reproduce the calculation steps [*Ceola et al.*, 2015; *Gil*, 2014; *Goodman et al.*, 2014].

A recent project (OntoSoft, <http://www.ontosoft.org/>) initiated the need for knowledge sharing about software development for geosciences, which attempts to promote an open research culture for geoscientists. In principal, geoscientists acknowledge and welcome the concept and benefit of open science [Kattge et al., 2014]. We found that though the “open science” concept for software development and data sharing may be discussed and practiced within a research community (e.g., Consortium of Universities for the Advancement of Hydrologic Science Inc., CUAHSI) involved in software development and data sharing [Rousseau et al., 2013], it is not routine to incorporate these practices in geoscience publication itself [Nosek et al., 2015]. Generally, the geoscience literature has little guidance for best practices for open science [Gil et al., 2016], and geoscientists prioritize their research motivations, discoveries, scientific questions, and engineering problems without emphasis on transparency and reusability of the research methods [Hsu et al., 2015]. As described by Gil et al. [2016], the latest review on the current state of the art on geoscience publication conducted suggested a strong interest in open, accessible, and reusable research, but feasible practices are needed to guide the implementation in each domain of geoscience.

The Penn State Hydrologic Model (PIHM), a coupled surface-subsurface model, is one of the models demonstrated in the OntoSoft project. Indeed, the “Geoscience Papers of the Future” (GPF) initiative in the OntoSoft project focuses on discussions about best practices in software and data sharing, provenance documentation, and digital scholarly publication. Here to digest and accommodate the GPF initiative, we adopt the tentative practice on software stewardship to promote research transparency and software utility for coupled surface-subsurface modeling studies. The goal of this study is to present typical guidelines and examples for hydrologic modelers on future scientific publication, so that the modeling simulation is transparent, reproducible, and reusable.

This paper builds on a brief background of PIHM and motivation of the reproducing study (section 2). We implemented GPF practices (section 3) based on the recommendations from GPF discussion [Gil et al., 2016]. Further, we demonstrated how a modeling study could be documented and reproduced according to the GPF practices (section 4). In section 5, we synthesized the users’ experiences from exploiters of the GPF practices. Then, we discussed the cost, benefit, and future implementation of GPF practices in section 6. The goal of the study is to enable transparency in the implementation of watershed hydrologic models that will also foster the reproducibility and reusability of coupled surface-subsurface models. The approach will also motivate activities to promote the lifetime and utility of software in geoscience publication. Finally, we expect that the implementing GPF best practices will stimulate a more sustainable approach for geoscience data and models.

2. Background

2.1. Knowledge Background of PIHM

The Penn State Integrated Hydrologic Model (PIHM) is a physics-based distributed coupled surface-subsurface hydrologic model (available online at <http://www.pihm.psu.edu/>). It simulates the terrestrial water cycle including interception, throughfall, snow accumulation and melt, infiltration, recharge, evapotranspiration, overland flow, variably saturated flow, groundwater flow, channel routing, and subchannel flow in a fully coupled scheme [Qu and Duffy, 2007; Kumar, 2009]. Evapotranspiration is calculated using the Penman-Monteith approach [Chen and Dudhia, 2001]. Overland flow is described with a 2-D estimation of St. Venant equations. Channel routing (i.e., streamflow) is modeled using a 1-D estimation of St. Venant equations. PIHM uses diffusive wave approximation for overland flow and streamflow. The model assumes that each subsurface layer can have both unsaturated and saturated storage components. Balance equations for both unsaturated and saturated zones are formed in a fully coupled way from Richard’s equation. Movement of moisture in unsaturated zones is assumed to be vertical only. The recharge to and from the water table is coupled with the moisture flux of unsaturated and saturated zones. For saturated groundwater flow, the 2-D Dupuit approximation is applied [Qu, 2005]. Interception and snowmelt are obtained using a Bucket Model [Qu, 2005] and a Temperature Index Model [Qu, 2005], respectively. Spatially, the modeling domain is decomposed into Delaunay triangles. The unstructured mesh allows users to resolve according to the geomorphological or hydrological characteristics of the watershed. The triangular mesh can be constrained by hydrological features, such as observation points or vector data (e.g., stream gauge, wells, soil maps, and land cover) and the watershed boundary conditions [Kumar, 2009]. The grid is projected vertically

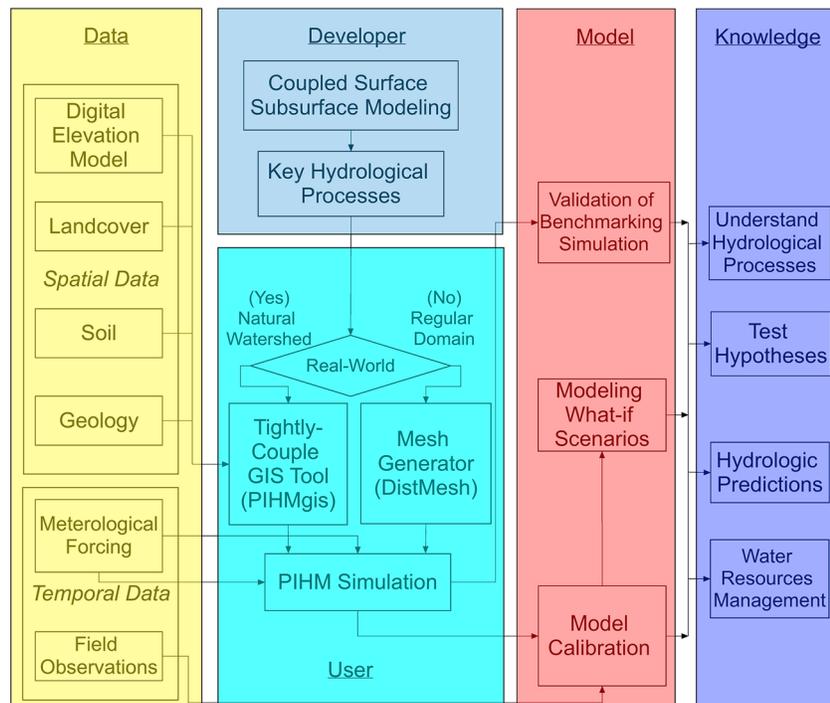


Figure 1. Application workflow of PIHM. PIHMgis is a GIS framework for distributed hydrologic modeling [Bhatt *et al.*, 2014]. DistMesh is a MATLAB code for generation of unstructured mesh [Persson and Strang, 2004].

to form either prismatic or trapezoidal/cuboidal volume elements on which physical equations will be formed afterward. The model resolves hydrological processes for evapotranspiration, overland flow, streamflow, and subsurface flow, governed by partial differential equations (PDEs). The PDE system is discretized on the projected elements 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 each model kernel. On each kernel, the representation of hydrological processes can be easily adjusted for certain research purposes. The flexible approach of coupling multiscale hydrological processes makes it adaptable for integrated hydrological simulation of diverse interests at different levels of complexity [Qu, 2005; Shi *et al.*, 2013; Duffy *et al.*, 2014].

Due to combined-mesh geometries (i.e., triangles and linear segments) used in model domain of PIHM, the ODEs are assembled as a stiff system. The system is then implemented with CVODE, a variable-order, variable-step, and multistep solver for stiff ODEs [Cohen and Hindmarsh, 1996]. Specifically, backward differentiation formulation and preconditioned Krylov method (generalized minimal residual method) are used to solve the large stiff system. Therefore, compilation of PIHM code requires the installation of Suite of Nonlinear and Differential/Algebraic Equation Solvers (SUNDIALS) [Hindmarsh *et al.*, 2005], which includes the CVODE package. In version 2.2 [Penn State Hydrologic Model (PIHM), 2015], PIHM started to publish both the executable files and source code, so that users do not have to worry about SUNDIALS installation to compile PIHM. Advanced users of PIHM can still modify the code and compile different versions after installation of SUNDIALS.

2.2. PIHM as a Basis for Community Models

The idea behind the development of PIHM was to support and realize the concept of “community models” for environmental predictions. After decades of development, the utility of PIHM spreads out to the areas of archeology [French and Duffy, 2014; Jazwa *et al.*, 2016], critical zone processes [Duffy *et al.*, 2014; Shi *et al.*, 2015b], flood assessment [Chen *et al.*, 2015], forest management [Yu *et al.*, 2015c], hydrologic data assimilation [Shi *et al.*, 2014, 2015a], landscape evolution [Zhang *et al.*, 2016], land surface energy balance [Shi *et al.*, 2013], scalable computing [Kumar and Duffy, 2015], and snowmelt runoff simulation [Kumar *et al.*, 2013; Wang *et al.*, 2013; Jepsen *et al.*, 2015]. A typical application workflow is shown in Figure 1. From the application workflow of PIHM,

who/what/when/where/why/how about research on PIHM are usually demonstrated, and the readers could decide if PIHM is appropriate for their research. The source code of PIHM was made freely available since the first version, as well as corresponding cyberinfrastructures [Yu *et al.*, 2015b] (e.g., data acquisition [Leonard and Duffy, 2013, 2014, 2016]). Therefore, we expected that PIHM would be learned and used independently by many others. Yet the evidence suggested that this has not been the case. The PIHM website showed 814 downloads worldwide since 2007. Nevertheless, reported studies on PIHM simulation were rather limited. We realized that open source has not lead to the expected utility of PIHM. This was probably because the application workflow only helps the readers to understand the research story. The computational processes were not clear in each modeling studies, which is crucial to the learning curve of PIHM.

2.3. Traditional Learning Curve of PIHM

Many users download PIHM to learn the source code line by line. Some users ask questions and report bugs by e-mailing the PIHM group members. As the number of users increased, the PIHM group organized three workshops, where attendees could gain hands-on experience. In addition, the group also hosted several visiting scholars individually to learn PIHM. The traditional learning process began with the geographic information system (GIS) data processing tool PIHMgis [Bhatt *et al.*, 2014]. Following these steps, which were recorded as video tutorials (<https://www.youtube.com/PIHMgis>), the information of a watershed was processed as the format of input files. Next, learners were recommended to understand the C source code of PIHM, compile on their own machine, and debug the code if there were any adaptations required. Finally, the learners could start to analyze PIHM simulation results to test their hypotheses.

Throughout the years, we found several common problems with the PIHM learning process. There were many adaptations and many versions of PIHM by different users for their own projects. However, in many publications of PIHM simulation, the authors only explained that the equations related to the adaptation, the details on code modifications, and developments were generally not provided. This could significantly affect the development and utility of PIHM. This was a common problem of many other hydrologic models. The adaptation was only explained as text in the publications. Sometimes, only a piece of pseudocode was provided. However, without explicit demonstration of the complete code, the adaptation could hardly be reused in later studies. Data preparation for PIHM requires learning of model source code due to complex data sets and parameters. PIHM would not be able to run smoothly due to a tiny mistake in data file format or bad combination of parameters. Therefore, users would have to debug the source code to locate the mistake. PIHM results were stored in plain txt files, which might cause confusion of units or meaning of hydrological variables. The analysis and visualization of model output required a lot of time and effort.

2.4. Problems in the Traditional PIHM Learning Example

A recent example of PIHM learning cost forced us to spend extra efforts to retrace and redo a finished project. The learning process took place with the distributed hydrologic modeling group at INRS (*Institut National de la Recherche Scientifique*), Canada. The group, Alain N. Rousseau, professor, Álvaro Pardo Álvarez, graduate intern, and Dominique Charron, undergraduate intern, wanted to learn PIHM to initiate an intercomparison project of hydrologic models [Rousseau *et al.*, 2015]. However, they were not able to rerun PIHM by solely relying on the paper of Qu and Duffy [2007] and other online tutorials and data sets. When they asked about details of the simulations and data processing pipelines, Xuan Yu had to look in backup files and redo the simulation to remind himself how the results were obtained. It was not easy because he had already graduated from the PIHM group at Penn State.

They first downloaded the input files of a finished PIHM project [Yu *et al.*, 2013]. Then, they spent considerable time on the compilation process of PIHM on their computers with the assistance of a computer technician. After installation of PIHM, they started to run PIHM and to understand the units and physical meaning of the input and output files. Due to a misunderstanding on the unit of the input files, they could not reproduce the simulation results shown in Yu *et al.* [2013]. As a consequence, they started to exchange e-mails with the PIHM group to figure out the irreproducible results. They had to set up several videoconference calls to discuss the misunderstandings and details of code and data. Finally, the simulation results were completely reproduced after around a half year.

This reproducibility study happened between two experienced hydrologic modeling groups, which motivated our focus on reusable simulation of hydrologic models. The INRS group has many experiences in

training and transferring hydrological modeling tools to researchers, students, and practicing water resources professionals [e.g., Fortin *et al.*, 2001; Noël *et al.*, 2014; Rousseau *et al.*, 2011; Turcotte *et al.*, 2001, 2003, 2007]. According to their learning procedure, we realized that what might be trivial for one modeling group may not be for the other. Such difference between groups may affect data sharing and reproducibility across the broad scientific community [Ceola *et al.*, 2015]. Only publishing reusable simulation of hydrologic models can effectively contribute to the water-related research community [Ceola *et al.*, 2015].

3. Implementing Best Practices: GPF

The GPF activity proposed best practices to make sure that geoscience papers of the future would capture the core concept of reproducibility [Gil *et al.*, 2016]. These best practices consist of 20 recommendations focusing on data accessibility, data documentation, software accessibility, software documentation, provenance documentation, method documentation, and author identification. We followed the recommendations listed in Gil *et al.* [2016] and explained the reason and our expectation of GPF practices in the context of our study.

3.1. Data Accessibility

Data accessibility remains a critical problem in the hydrologic literature. Often, access to data sets is not provided in the journal article. Typically, there are references cited to clarify sources of data sets. However, in watershed models, the data are heterogeneous and reside on many national data servers [Leonard and Duffy, 2013], and readers have to prepare the data set again according to the references or the readers find guidelines by contacting the corresponding author or agency via e-mails. However, e-mails have limited attachment size and require response time according to the authors' availability. A recent study demonstrated that less than half of authors are willing to share data upon request [Krawczyk and Reuben, 2012]. Therefore, a culture of making data sets public should be nurtured to guarantee accessibility and reusability. Here we choose Figshare (<http://figshare.com>) as the online digital repository to make all the data related to the study accessible. The data sets were licensed under CC0 (<https://creativecommons.org/about/cc0>) and were cited in the papers and persistently accessible to the readers.

Data accessibility is a key step for reproducibility, especially for the intermediate data. Public repository with a permanent digital object identifier (DOI) is the best way to make data accessible. In hydrologic modeling studies, we usually provide a table listing all the initial conditions, boundary conditions, forcing variables, and parameters along with units. Even in the simulation of benchmarking cases [Maxwell *et al.*, 2014], the actual input files vary dramatically among models and modelers' habits. Therefore, it is more efficient and reusable to explicitly provide the access of these data files. Besides figshare, there are other repositories to upload data and obtain a DOI, such as Zenodo, Dryad, and ResearchGate.

3.2. Data Documentation

Data documentation is to establish the metadata. General-purpose metadata usually includes the data set title and description, format, author, and date created to make finding and utilizing the data easier. Notably, the physical terms used in metadata may be different according to disciplinary backgrounds and customs. Therefore, it is recommended to think about the background of potential readership. The metadata should be accurate.

Typically, hydrologic model input/output data require documentation of format, unit, and spatial and temporal scale, so that mathematical variables can be related to hydrologic parameters, storages, and fluxes. For example, Mirus *et al.* [2011] documented a data set to support future community use. PIHM provides general documentation about input and output data format. In each specific modeling study of PIHM, we think that data documentation is still necessary because the readers may want to figure out how to reuse the data files and how to find similar data that suit their purposes.

3.3. Software Accessibility

Software is used to process initial data and to generate any intermediate or final results and figures. PIHM is available on GitHub (<https://github.com/>), with a license and a DOI. In the GitHub repository, both source code and executable files are provided. The DOI should be cited in the paper to make sure that the readers could always obtain the exact version of model that was used. Besides the PIHM code, we also provided the

postprocessing codes as part of the software used in each study. These parts of software were important for helping readers understand model results.

There are some other coupled surface-subsurface models hosted on public repositories. The OpenGeoSys (OGS) [Kolditz *et al.*, 2012] community has been using GitHub (<https://github.com/ufz/ogs>) since 2010. Their experiences demonstrated that making software accessible is a beneficial activity for the hydrologic modeling community. The PFLOTRAN [Lichtner *et al.*, 2013] community uses Bitbucket (<https://bitbucket.org/pflotran/>) to realize the collaborative code development.

3.4. Software Documentation

Software documentation helps the readers understand the utility of the software. We have been asked many times: do you think PIHM is the ideal model for my project, how to create similar plots in your paper, how did you prepare the input files, and where is the code to manipulate the output files? Software documentation will answer such questions about all the codes used in the research paper.

PIHM is described in the OntoSoft portal (<http://www.ontosoft.org/portal>). OntoSoft relies on ontology to capture scientific software metadata [Gil *et al.*, 2015]. The metadata captured by OntoSoft include six major categories based on information that a scientist would seek about the software: (1) identify software, (2) understand and assess software, (3) execute software, (4) get support for the software, (5) do research with the software, and (6) update the software. Such information allows the users to find the software that is relevant to them.

PIHM is registered in Community Surface Dynamics Modeling System (CSDMS) model repository, where 197 geoscience models are listed. In the repository, PIHM is described with the model metadata including technical specification, input/output, and physical process description. The variable names of PIHM are mapped to a standard name database [Peckham, 2014] to accelerate cross-discipline collaboration. The mapping result of PIHM variables to CSDMS standard name is listed in Supporting Information S1. Such information allows users from different background to understand the physical meaning of each variable in PIHM.

Besides PIHM, there is other important software indispensable to the research, including codes used to manipulate data, process results, and plot figures. Since these pieces of software consist of a few lines of code, we use simple description to document them. Such software can help the readers understand computational workflow and precisely repeat the whole research.

3.5. Provenance Documentation

Here provenance means the origin or source of the data used and the processing performed to reach a scientific result [Moreau, 2010]. Usually, we use a computational workflow to capture the provenance. Computational workflow (or scientific workflow) is the information which scientists can model, design, execute, and repeat their analysis and visualization pipelines [Barga and Digiampietri, 2008]. Notably, the computational workflow is different from the application workflow, which generally describes the sequence of tasks in a project (e.g., Figure 1).

In the studied cases, we started from the original data sets and software and then described how they were performed, what intermediate data sets were produced, and how figures are plotted to support our arguments and ideas. The computational workflow is the key information to guarantee the reproducibility.

3.6. Method Documentation

Method documentation is a generalized version of provenance, which may provide insights in readers' future applications. The method documentation helps the readers to decide and predict if the applied modeling strategy is appropriate to their data and hypotheses. Usually, the traditional way of method section in the article is sufficient for the readers to understand the modeling strategy.

3.7. Author Credit and Identification

Author identification clarifies the role and contribution of each author, which helps the reader better understand the collaboration process. Author identification also provides the persistent information of each author so that readers can easily find the contact of the author for questions.

Table 1. Author Identifiers and Contributions

Author Name	ORCID	Contributions
Xuan Yu	0000-0002-9055-7923	Led the conception and design of the research, implemented GPF practices, and wrote the paper.
Christopher J. Duffy	0000-0003-0080-6445	Led the conception and design of the research and revised the paper.
Alain N. Rousseau	0000-0002-3439-2124	Led the conception and design of the research and revised the paper.
Gopal Bhatt	0000-0002-6627-793X	Prepared the data and software and revised the paper.
Álvaro Pardo Álvarez	0000-0003-0302-9856	Provided end user experiences and revised the paper.
Dominique Charron	0000-0002-8494-1152	Provided end user experiences and revised the paper.

We used Open Researcher and Contributor ID (ORCID) [Haak et al., 2012] to persistently identify the authors for this study. The author information and contributions are listed in Table 1.

4. Cases Studies of Documenting PIHM Simulations

We chose two examples to implement GPF. The first example (Case A) is the V-shaped catchment, which is a synthetic problem on a regular domain [Di Giammarco et al., 1996]. The second example (Case B) is a real-world hydrologic application on a headwater catchment [Qu and Duffy, 2007]. The PIHM simulations started from PIHM input files to result visualization. Please note that setting up a PIHM simulation requires preprocessing of domain discretization [e.g., Persson and Strang, 2004] and geospatial data integration [Bhatt et al., 2014]. Case A used a triangular mesh (782 triangles) to represent the model domain. Case B used PIHMgis [Bhatt et al., 2014] to create a triangular mesh of the watershed. Once the input files are prepared, we started to document PIHM simulation.

4.1. Workflow of PIHM Simulation

The computation workflow (Figure 2) started with the input files accessible from the figshare repository [Yu, 2015a, 2015b, 2016a]. The PIHM v2.2 is accessible at the GitHub repository [PIHM, 2015]. We used mesh3dplot.m [Yu, 2016b] to visualize the physical representation of the watershed. The PIHM simulation

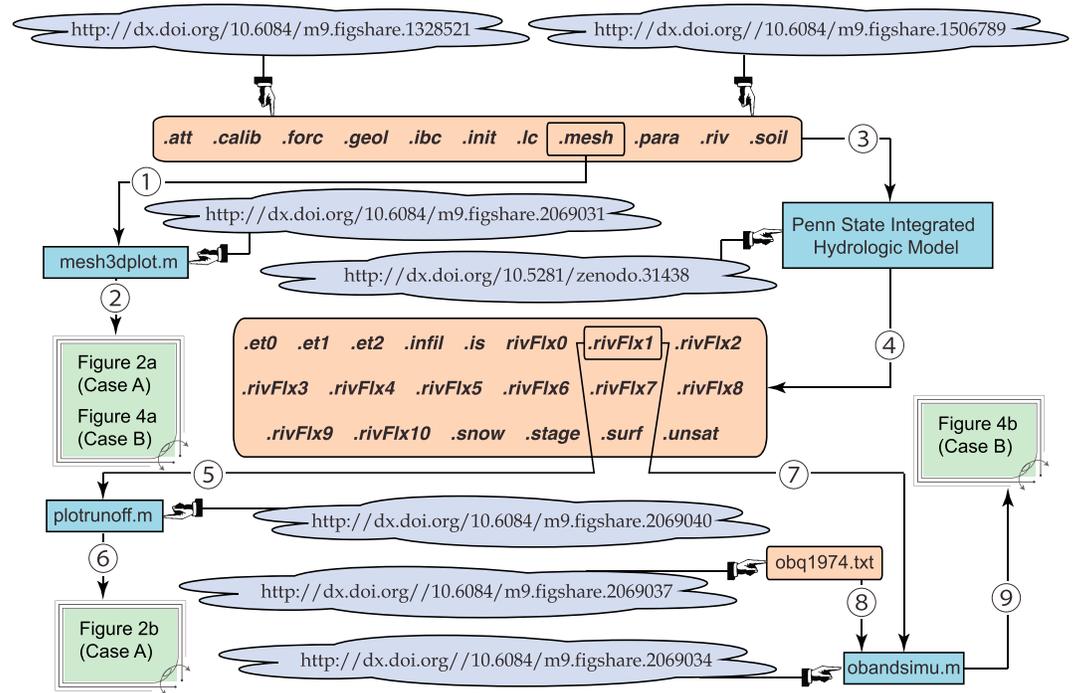


Figure 2. The computational workflow of PIHM case studies. The flowchart starts with data (in orange) and software (in blue) from repositories (indicated by a small hand) and ends with the figures in the publication (in green). The numbers indicate the order of the data processing.

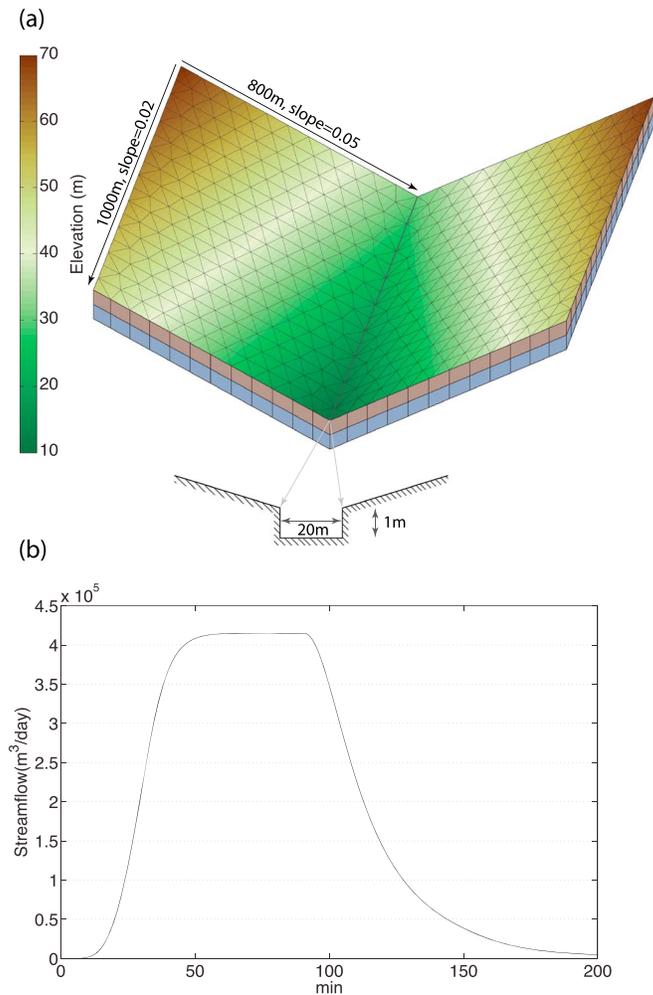


Figure 3. PIHM simulation at V-shaped catchment. (a) The domain geometry and mesh. (b) Modeled runoff response of a 90 min duration and 3×10^{-6} m/s intensity rainfall event.

generated a set of the plain txt files. We used the postprocessing script [Yu, 2016c, 2016d] to plot the runoff at the outlet (Figures 3b and 4b).

4.2. Case A: V-Shaped Catchment

The V-shaped catchment has been widely applied to verify 2-D overland surface flow coupled with channel routing [Di Giammarco et al., 1996]. In the example, overland flow from a tilted V-shaped catchment was generated by a rainfall event of 90 min duration and 3×10^{-6} m/s intensity. The dimension and model domain of the V-shaped catchment are shown in Figure 3a. The domain was made of two 1000 m by 800 m planes connected to a channel of 1000 m long and 20 m wide. Land surface slopes were 0.05 and 0.02, perpendicular to and parallel to the channel, respectively. Manning roughness was 0.015 for the sloping land surface and 0.15 for the channel. The boundary conditions were no flow at the plane and critical depth at the outlet. The final plot of the runoff suggested that it took around 60 min for the outlet runoff to reach the maximum flow rate. The simulation results were consistent with those produced by other coupled surface-subsurface models [Maxwell et al., 2014].

4.3. Case B: Shale Hills Watershed

In 1974, a field experiment was conducted at the Shale Hills Watershed (Figure 4a) in central Pennsylvania to observe the role of antecedent moisture (initial conditions) on runoff response using artificial rainfall events. Qu and Duffy [2007] developed the first version of PIHM to simulate the observed runoff and groundwater

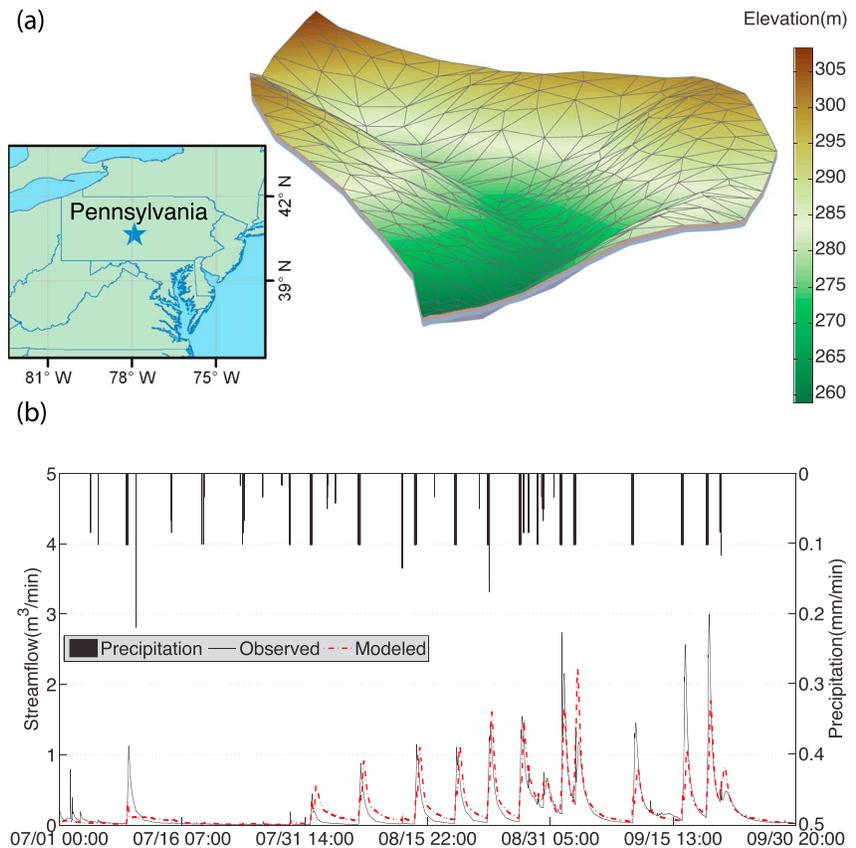


Figure 4. PIHM simulation at Shale Hills Watershed. (a) Watershed geometry and mesh. (b) Observed and modeled streamflow during the 1974 irrigation experiment.

levels. Here we reproduced the study with PIHMv2.2. The decomposition of the watershed is shown in Figure 4a. From the final simulation (Figure 4b), we could see the same amount of rainfall generating different runoff responses in both observed and modeled streamflow. The reproduced experiment here only focused on the runoff response. More details about soil moisture and groundwater dynamics during the 1974 irrigation experiment could be found in *Qu and Duffy [2007]*.

5. Testing the User Experience With GPF Practices

The goal of this section is to test if the user/reader is able to completely reproduce the PIHM simulations following these GPF practices. Here we asked Álvaro Pardo Álvarez (a beginner who had never heard about PIHM before and then an intermediate user of PIHM) and Dominique Charron (a beginner who had never heard about PIHM before) to provide user experiences. Specifically, we had the following expectations: (1) be able to download the PIHM source code and compile it on their computer or directly download the executable file of PIHM, (2) be able to download the input data for the two case study examples of PIHM, (3) be able to run the PIHM simulation, (4) be able to download the data analysis code to generate the figures, (5) improve understanding of hydrological processes by following the computational workflow, and (6) create a standardized and reusable implementation strategy in their future project reports and scientific papers.

5.1. Users' Assessment

GPF practices significantly reduced the time cost of reproducing the computational steps (how) and understanding the corresponding hydrological meaning (why). The assessment is based on users' reproducing processes of the computational workflow shown in Figure 2. We recorded the time each user spent and their learning curve of model-related hydrologic knowledge (Table 2). From the time spent, we can tell that Álvaro

Table 2. Time Cost of Reproducing Computational and Hydrologic Processes

Step in Workflow (Figure 2)	How?—Time Cost of Reproducing Computational Processes (Minute)					Why?—Time Cost of Understanding Hydrologic Processes (Minute)				
	Beginner			Intermediate		Beginner			Intermediate	
	ÁP		DC	ÁP		ÁP		DC	ÁP	
	No GPF	GPF	GPF	No GPF	GPF	No GPF	GPF	GPF	No GPF	GPF
0	60	5	~5	30	5	2400	960	~5	960	120
1	40	2	~2	20	2	30	2	~2	10	2
2	5	2	~2	3	2	15	2	~2	7	2
3	5	3	~15	3	3	15	5	~10	5	3
4	90	40	~30	40	30	30	15	~15	15	8
5	30	3	~5	15	3	20	2	~2	10	2
6	5	3	~5	3	2	10	2	~2	5	2
7	30	3	~5	15	2	20	2	~2	10	2
8	5	2	~2	5	2	10	2	~5	5	2
9	5	2	~2	5	2	10	2	~2	5	2

Pardo Álvarez and Dominique Charron were able to reproduce the computational workflow in around an hour when GPF was presented. GPF also improved the efficiency to understand the hydrologic processes. Their evaluation suggested that all the GPF practices are useful (Table 3). They would like to try GPF in their own future publications, though the time spent for preparation of GPF is significant. Notably, the time spent (or cost) for the preparation of GPF will reduce when the authors get used to it.

Time cost of understanding hydrologic processes suggested subjective differences in the reproducing processes. When two users are given the same data set and computational workflow, the time cost for understanding hydrologic processes is significantly different. Surprisingly, the beginner (Dominique Charron) used much less time to understand the hydrologic processes than an intermediate user (Álvaro Pardo Álvarez). This is probably due to the subjective difference in the levels of understanding. A control on subjective difference will be necessary for the reproducibility, understanding, and sharing of hydrologic knowledge [Ceola et al., 2015].

5.2. Álvaro Pardo Álvarez’s Comments

Based on my experience using PIHM, I can confirm the reliability of the runoff results described in the two PIHM simulations. I downloaded the PIHM executable files. I also downloaded the input files for the two examples. I ran the PIHM simulation on my windows machine. The downloaded data analysis codes were able to reproduce Figures 3 and 4.

Table 3. Survey on Usefulness of GPF

GPF	Usefulness on a Scale From 0 to 10	
	ÁP	DC
Data accessibility	10	10
Data documentation	9	8
Software accessibility	9	9
Software documentation	8	8
Provenance documentation	10	10
Method documentation	9	10
Author identification	10	9
Time cost estimation of GPF practice in your own papers (days)		
4 (ÁP)		7 (DC)
Are you still likely going to implement GPF practice in your own papers (on a scale from 0 to 10)?		
9 (ÁP)		9 (DC)

I believe that GPF activity will highly help PIHM users, not just beginners but also advanced users, to conduct and reproduce simulations without losing time on input data preparation and output data visualization. Additionally, it will give them a quick start on learning coupled surface-subsurface modeling thanks to the description of the modeling details. All these facts make this type of activities extremely helpful to the geoscience community by sharing knowledge, promoting transparency in research, and, ultimately, stimulating some new thinking in open

science. Based on this, I intend to implement similar activities in future prospective studies (research papers or scientific/technical reports).

5.3. Dominique Charron's Comments

Following the information included in a draft version of this paper and the links in the workflow (Figure 2), I was able to download the model and run the specific simulation of the V catchment and the Shale Hills. The workflow explained how to set the directories to run the models, and with the MATLAB files packaged with the download I was able to get the same results as those presented in the article (Figures 3 and 4). I was able to achieve that within a few hours. In this regard, the procedure listed is efficient and provides reproducible results. I am interested to implement such practices for my future studies.

5.4. Alain N. Rousseau's Comments

As mentioned earlier in this paper, we can qualify our group as an experienced group in transferring and training hydrological modeling tools to researchers, students, and practicing water resources professionals [Rousseau *et al.*, 2013]. Our supporting documents include users' manuals, technical reports and peer-reviewed papers, and last but not least a community of end users. That is why we think that this initiative of embedding open science practice in each paper represents an insightful contribution to the hydrological modeling community as well as the geoscience community at large.

6. Discussion

GPF best practices serve not only for learning a particular model but also for any research paper with a data analysis emphasis, though we use the learning curve of the model PIHM as the example here. Of course, the practicality of implementation of best practices within the broader geoscience community must be balanced by the effort or cost, benefits, and rewards of GPF practices before fostering them as routine.

6.1. Cost of GPF Practices

The degree of effort to implementation of GPF practices involves learning digital scholarship in geosciences. In general, GPF practices are not currently required in most geoscience journals. GPF practices require extra work for the authors. For example, the source code needs to be programmed with self-explained (i.e., meaningful) variable and parameter names, faster loops [Easterbrook, 2014]. A lot of code notes should be prepared. Sometimes, the program might be validated in different computational environments. Each data file requires corresponding documentation. The provenance should be presented with a workflow of all the data sets, processing steps, and plotting methods. Clearly, the authors must value best practices if they are going to invest the time to implement them in their manuscripts.

6.2. Benefit of GPF Practices

Documenting the benefits of GPF practices has both long-term and short-term implications. A short-term reward for authors is that their research is more accessible, easier to understand the methods and results, and, thus, is more likely to be cited or their model more likely to become a community resource. The community will be able to efficiently learn the workflow from data, to model simulation, to results without the uncertainty that their interpretation of the model-data structure is incorrect. Code reuse allows the future young scientist to test even the smallest assumptions in a physical process and the underlying code.

The GPF practices will benefit authors in other ways beyond data and software reuse and citations. The details of the computational workflow will be published and will eliminate trivial author communication to explain access and computational details to potential users. It is often the case that good documentation, in this case for the data and the model, leads to community adoption simply because the workflow is clear and the results are understandable. The authors will not have to backup any model simulations after it is published, because all the code and data are accessible from the paper. In the future, the community adopting best practices will improve the coding quality and collectively advance the fidelity, reliability, and efficiency of the model.

The GPF practices will be beneficial to the peer review process. Increasing collaborative papers are emerging with complex knowledge across disciplines, which increase the difficulty for editors to find reviewers and make reviewers work harder to evaluate the papers [Hayden, 2015]. GPF practices demonstrate the

value of efficient reproducibility, so that even multidisciplinary reviews are possible [Nekrutenko and Taylor, 2012].

6.3. A Call for Routine GPF Practices in Publication

We expect that the GPF practices will make substantial contribution to the development of coupled surface-subsurface models, even though it will require time and effort beyond preparing the manuscript. Of course, there are many ways to promote the reproducibility and reusability of coupled surface-subsurface models. For example, from the experience of this limited study of applying GPF practices for complex hydrologic model and data, it is clear to us that best practices recommended by *Gil et al.* [2016] can be reasonably implemented as a component of publication. Given the complexity of the data and model used here, the experiment clearly has application to the broader geosciences, and we encourage a broad debate for promoting lifetime and utility of Earth science software and data. As a start to this conversation, we summarize the PLUS guidelines for geoscience papers to make the software and data persistent, linked, user-friendly, and sustainable (Figure 5). *Persistent*. Data, software, and authors should be persistently (i.e., consistently) identifiable. The traditional publications have DOIs (digital object identifiers) after the proofreading. Articles can be easily accessible through those identifiers. In GPF practices, the process of the study (including data and software) should be also assigned with persistent unique identifiers (e.g., data and software can be uploaded to online repositories with DOIs or PURLs (Permanent Uniform Resource Identifiers)). *Linked*. Data and software should be linked in the computational workflow so that the software can be understood and reused by the readers. These links should include any intermediate data derived from the original data that represent essential information in the final figures of the article. *User-friendly*. To make software user-friendly is useful for the discovery and reuse among readers. Metadata are always good for the users of the data and software. The software should be packaged with documentation and instructions so that readers can first decide if the software can be used in their work, and how to apply it. It is always good to be thoughtful for any possible readership, especially those from other disciplines. *Sustainable*. Authors are recommended to register an ORCID (Open Researcher and Contributor ID), so that readers can always obtain the research updates. Software should be maintained at repositories (e.g., GitHub, CRAN, and CodePlex) so that further development can be achieved (e.g., users will be notified when the software is updated and authors will receive suggestions and comments regarding further development). Software license should have instructions on copy, modification, redistribution, and development.

6.4. Implications for GPF Best Practices

GPF best practices will improve reproducibility of papers. Reproducibility can be achieved by rerunning the same code, or by writing a new program, or by designing a new computational experiment [Easterbrook, 2014]. GPF directly preserves the computational workflow and targets for reproducibility by rerunning the same code. *Ceola et al.* [2015] suggested that a detailed protocol is crucial to ensure the comparability and reproducibility of experiment across groups. Such protocol controls the errors (e.g., subjective preference, domain-specific terminology [Peckham, 2014], and group dependent practice) in each intermediate data set and steps. GPF preserves all the intermediate data and code, which will definitely facilitate scientific reproducibility by a new program or a new computational experiment.

GPF best practices are important to illumination of dark data. Pairing data along with papers is an efficient way to uncover the dark data [Gil and Hirsh, 2012]. From the text part of a paper, readers can be attracted by the story, which arouses their curiosity of the data part. Since all the data, software, and figures are linked, the readers will be able to understand and uncover the digital scholarship based on the science story.

GPF best practices will improve the understanding of multidisciplinary processes. Earth scientists usually use tools from physics, chemistry, biology, chronology, and mathematics to improve the understanding of processes with diverse temporal and spatial scales. It turns out to be many subdisciplines with limited cross-disciplinary actions. Some multidisciplinary research is just integration of different pieces of knowledge and stories. The big challenge is to unravel the processes behind the different subdisciplines [Acocella, 2015]. Clearly, GPF will accelerate the understanding process across disciplines by explicitly linking raw data, research methods, and science stories.

GPF best practices will benefit the dissemination and communication to the society of the results, achievements, and general outcome of the research in geoscience. Human activities are now a dominant geological

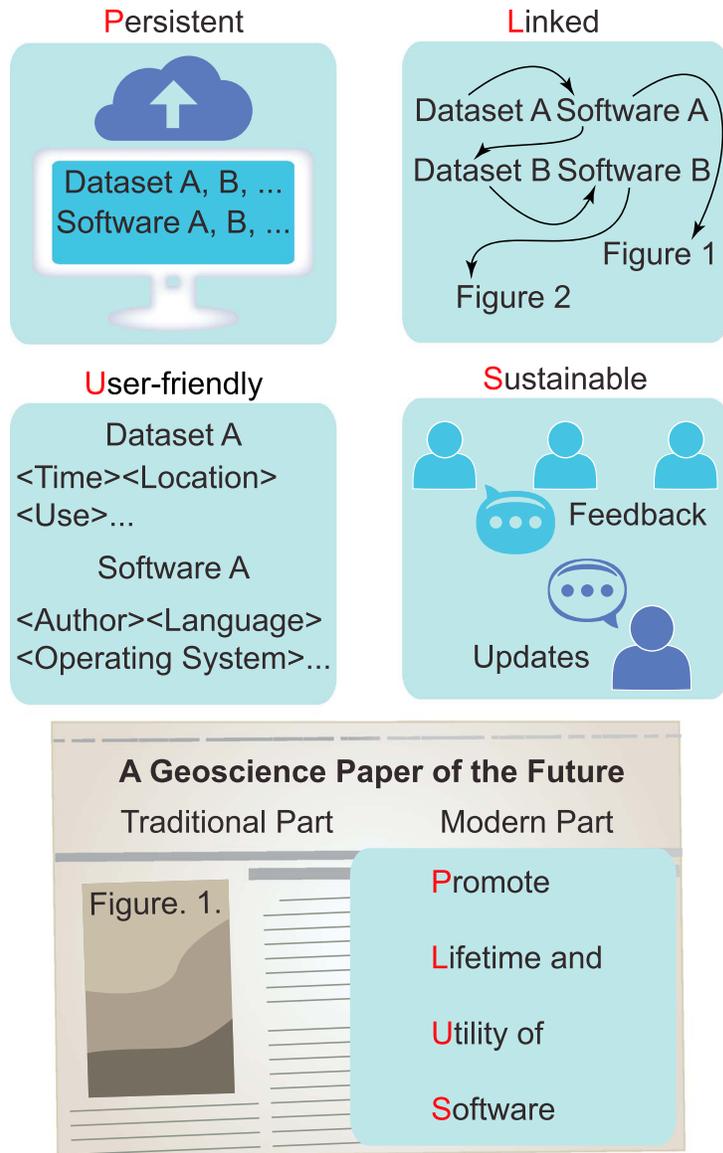


Figure 5. The “PLUS” guideline for geoscience papers of the future. The PLUS guideline tries to make data and software persistent, linked, user-friendly, and sustainable. Future geoscience papers should include such a modern part to promote lifetime and utility of software.

force on the Earth; therefore, public communication and societal engagement is an important aspect of geoscience research [Stewart, 2016; Buytaert et al., 2014]. GPF improves research transparency, which provides valuable information for decision makers and build public confidence in supporting geosciences [Acocella, 2015].

7. Summary and Conclusion

In this study, we implemented the GPF practices to improve the transparency of coupled surface-subsurface modeling procedures and results and evaluated the effectiveness of these practices. The users, which spanned a wide range of experience, agreed that reproducibility of the model led to a deeper understanding of the model physics and the supporting data. Furthermore, given the range of experience of the users, it provided important feedback to the developers. This feedback is critical for advancing community models. We also foresee that adopting the GPF practices will increase the fidelity of simulation results, reduce the learning

curve, and enhance the model utility. Specifically, the example of PIHM will realize its concept of community models with a broad range of users including teachers, students, professionals, and the educated public.

While there are still many questions left unanswered about reusability of model simulations in this experiment, we have aimed to establish one concept: the traditional way of geoscience publication does not lead to or support the reuse of data and software. Moreover, we attempted to test the role of digital scholarship in promoting an open research community culture, aiming at model development, model implementation studies, and coordinated field and laboratory data analyses. The way forward is to develop clear, community guidelines and consensus that support best practices for software and data reusability. This can only happen as such practices are proposed, tested, and adopted by the research community and supported by the funding agencies and publication policies [Nosek *et al.*, 2015; McNutt *et al.*, 2016], such as the U.S. National Science Foundation (NSF) Public Access Plan in March 2015 [National Science Foundation, 2015]; the U.S. Geological Survey (USGS) issued a Data Management Policy to ensure that the results of scientific research are available and reusable [U.S. Geological Survey, 2015]; *Journal of Visualized Experiments* (<http://www.jove.com/>) was established in 2006 to improve research reproducibility and reduce learning curve; *Vadose Zone Journal* launched Reproducible Research (RR) program to guide the data and code publication alongside the articles [Skaggs *et al.*, 2015]; editors of several hydrologic journals decided to improve the peer review process toward enhanced transparency [Koutsoyiannis *et al.*, 2016]. Shaping new digital scholarship and disciplinary norms is an important step to a fair and open research culture in geoscience.

Acknowledgments

All data for this paper are properly cited and referred to in the reference list. More details of the data are available upon request from the corresponding author (xuan@udel.edu). This work was supported by NSF grants IIA-1301765, IIS-1344272, ICER-1343800, and ICER-1440323 and Shale Hills-Susquehanna Critical Zone Observatory (EAR-0725019, EAR-1239285, and EAR-1331726). The authors thank the EarthCube OntoSoft leadership for initiating the Geoscience Papers of the Future program and the early career advisory committee members. The manuscript benefited from conversations facilitated by the EarthCube Distinguished Lecturer Program. The authors also thank Yolanda Gil of University of Southern California for the insightful discussions and comments on the manuscript. We also want to thank Ibrahim Demir from University of Iowa and Runjian Wu from University of Arizona for their help with proofreading. We are very grateful for the constructive comments from reviewer Christopher Hutton and an anonymous reviewer on the earlier versions of the manuscript.

References

- Acocella, V. (2015), Grand challenges in Earth science: Research toward a sustainable environment, *Front. Earth Sci.*, *3*, 68, doi:10.3389/feart.2015.00068.
- Barga, R. S., and L. A. Digiampietri (2008), Automatic capture and efficient storage of e-Science experiment provenance, *Concurrency Comput.: Pract. Exper.*, *20*, 419–429, doi:10.1002/cpe.1235.
- Bhatt, G., M. Kumar, and C. J. Duffy (2014), A tightly coupled GIS and distributed hydrologic modeling framework, *Environ. Modell. Software*, *62*, 70–84, doi:10.1016/j.envsoft.2014.08.003.
- Brooks, P. D., J. Chorover, Y. Fan, S. E. Godsey, R. M. Maxwell, J. P. McNamara, and C. Tague (2015), Hydrological partitioning in the critical zone: Recent advances and opportunities for developing transferable understanding of water cycle dynamics, *Water Resour. Res.*, *51*, 6973–6987, doi:10.1002/2015WR017039.
- Buytaert, W., Z. Zulkafli, S. Grainger, L. Acosta, T. C. Alemie, J. Bastiaensen, B. De Bièvre, J. Bhusal, J. Clark, and A. Dewulf (2014), Citizen science in hydrology and water resources: Opportunities for knowledge generation, ecosystem service management, and sustainable development, *Front. Earth Sci.*, *2*, 26, doi:10.3389/feart.2014.00026.
- Ceola, S., B. Arheimer, E. Baratti, G. Blöschl, R. Capell, A. Castellarin, J. Freer, D. Han, M. Hrachowitz, and Y. Hunechea (2015), Virtual laboratories: New opportunities for collaborative water science, *Hydrol. Earth Syst. Sci.*, *19*(4), 2101–2117, doi:10.5194/hess-19-2101-2015.
- Chen, F., and J. Dudhia (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, doi:10.1175/1520-0493(2001)129<0569:CAALSH>2.0.CO;2.
- Chen, X., M. Kumar, and B. L. McGlynn (2015), Variations in streamflow response to large hurricane-season storms in a southeastern U.S. watershed, *J. Hydrometeorol.*, *16*(1), 55–69, doi:10.1175/JHM-D-14-0044.1.
- Clark, M. P., Y. Fan, D. M. Lawrence, J. C. Adam, D. Bolster, D. J. Gochis, R. P. Hooper, M. Kumar, L. R. Leung, and D. S. Mackay (2015), Improving the representation of hydrologic processes in Earth System Models, *Water Resour. Res.*, *51*, 5929–5956, doi:10.1002/2015WR017096.
- Cohen, S. D., and A. C. Hindmarsh (1996), CVODE: A stiff/nonstiff ODE solver in C, *Comput. Phys.*, *10*(2), 138–143.
- De Schepper, G., R. Therrien, J. C. Refsgaard, and A. L. Hansen (2015), Simulating coupled surface and subsurface water flow in a tile-drained agricultural catchment, *J. Hydrol.*, *521*, 374–388, doi:10.1016/j.jhydrol.2014.12.035.
- Dickerson, D., and K. Dawkins (2004), Eighth grade students' understandings of groundwater, *J. Geosci. Educ.*, *52*(2), 178–181.
- Di Giammarco, P., E. Todini, and P. Lamberti (1996), A conservative finite elements approach to overland flow: The control volume finite element formulation, *J. Hydrol.*, *175*(1), 267–291, doi:10.1016/S0022-1694(96)80014-X.
- Duffy, C., Y. Shi, K. Davis, R. Slingerland, L. Li, P. L. Sullivan, Y. Goddérís, and S. L. Brantley (2014), Designing a suite of models to explore critical zone function, *Procedia Earth Planet. Sci.*, *10*, 7–15, doi:10.1016/j.proeps.2014.08.003.
- Easterbrook, S. M. (2014), Open code for open science?, *Nat. Geosci.*, *7*(11), 779–781, doi:10.1038/ngeo2283.
- Fatichi, S., et al. (2016), An overview of current applications, challenges, and future trends in distributed process-based models in hydrology, *J. Hydrol.*, *537*, 45–60, doi:10.1016/j.jhydrol.2016.03.026.
- Fortin, J., R. Turcotte, S. Massicotte, R. Moussa, J. Fitzback, and J. P. Villeneuve (2001), Distributed watershed model compatible with remote sensing and GIS data. I: Description of model, *J. Hydrol. Eng.*, *6*(2), 91–99, doi:10.1061/(ASCE)1084-0699(2001)6:2(91).
- Fossey, M., A. N. Rousseau, F. Bensalma, S. Savary, and A. Royer (2015), Integrating isolated and riparian wetland modules in the PHYSITEL/HYDROTEL modelling platform: Model performance and diagnosis, *Hydrol. Processes*, *29*, 4683–4702, doi:10.1002/hyp.10534.
- French, K. D., and C. J. Duffy (2014), Understanding ancient Maya water resources and the implications for a more sustainable future, *Wiley Interdiscip. Rev. Water*, *1*(3), 305–313, doi:10.1002/wat2.1024.
- Gil, Y. (2014), *Intelligent Workflow Systems and Provenance-Aware Software Proceedings of the 7th International Congress on Environmental Modelling and Software*, edited by D. P. Ames, N. W. T. Quinn, and A. E. Rizzoli, San Diego, Calif., 15–19 Jun.
- Gil, Y., and H. Hirsh (2012), Discovery informatics: AI opportunities in scientific discovery, paper presented at AAAI Fall Symposium on Discovery Informatics: The Role of AI Research in Innovating Scientific Processes, Arlington, Va.
- Gil, Y., V. Ratnakar, and D. Garijo (2015), OntoSoft: Capturing scientific software metadata, in *Proceedings of the 8th ACM International Conference on Knowledge Capture, Palisades, NY, USA*, pp. 1–4, ACM, New York.

- Gil, Y., et al. (2016), Towards the geoscience paper of the future: Best practices for documenting and sharing research from data to software to provenance, *Earth Space Sci*, doi:10.1002/2015EA000136.
- Golden, H. E., C. R. Lane, D. M. Amatya, K. W. Bandilla, H. R. Kiperwas, C. D. Knightes, and H. Ssegane (2014), Hydrologic connectivity between geographically isolated wetlands and surface water systems: A review of select modeling methods, *Environ. Modell. Software*, *53*, 190–206, doi:10.1016/j.envsoft.2013.12.004.
- Goodman, A., A. Pepe, A. W. Blocker, C. L. Borgman, K. Cranmer, M. Crosas, R. Di Stefano, Y. Gil, P. Groth, and M. Hedstrom (2014), Ten simple rules for the care and feeding of scientific data, *PLoS Comput. Biol.*, *10*(4), e1003542, doi:10.1371/journal.pcbi.1003542.
- Haak, L. L., M. Fenner, L. Paglione, E. Pentz, and H. Ratner (2012), ORCID: A system to uniquely identify researchers, *Learned Publishing*, *25*(4), 259–264, doi:10.1087/20120404.
- Hayden, E. C. (2015), Rule rewrite aims to clean up scientific software, *Nature*, *520*, 276–277, doi:10.1038/520276a.
- Heidorn, P. B. (2008), Shedding light on the dark data in the long tail of science, *Library Trends*, *57*, 280–99.
- Hindmarsh, A. C., P. N. Brown, K. E. Grant, S. L. Lee, R. Serban, D. E. Shumaker, and C. S. Woodward (2005), SUNDIALS: Suite of Nonlinear and Differential/Algebraic Equation Solvers, *ACM Trans. Math. Software*, *31*(3), 363–396.
- Hsu, L., R. L. Martin, B. McElroy, K. Litwin-Miller, and W. Kim (2015), Data management, sharing, and reuse in experimental geomorphology: Challenges, strategies, and scientific opportunities, *Geomorphology*, *244*, 180–189, doi:10.1016/j.geomorph.2015.03.039.
- Jazwa, C. S., C. J. Duffy, L. Leonard, and D. J. Kennett (2016), Hydrological modeling and prehistoric settlement on Santa Rosa Island, California, USA, *Geoarchaeology*, doi:10.1002/gea.21532.
- Jepsen, S. M., T. C. Harmon, M. W. Meadows, and C. T. Hunsaker (2015), Hydrogeologic influence on changes in snowmelt runoff with climate warming: Numerical experiments on a mid-elevation catchment in the Sierra Nevada, USA, *J. Hydrol.*, doi:10.1016/j.jhydrol.2015.12.010, in press.
- Kattge, J., S. Diaz, and C. Wirth (2014), Of carrots and sticks, *Nat. Geosci.*, *7*(11), 778–779, doi:10.1038/ngeo2280.
- Kolditz, O., S. Bauer, L. Bilke, N. Böttcher, J. O. Delfs, T. Fischer, U. J. Görke, T. Kalbacher, G. Kosakowski, and C. I. McDermott (2012), OpenGeoSys: An open-source initiative for numerical simulation of thermo-hydro-mechanical/chemical (THM/C) processes in porous media, *Environ. Earth Sci.*, *67*(2), 589–599, doi:10.1007/s12665-012-1546-x.
- Koutsoyiannis, D., G. Blöschl, A. Bárdossy, C. Cudennec, D. Hughes, A. Montanari, I. Neuweiler, and H. Savenije (2016), Joint editorial—Fostering innovation and improving impact assessment for journal publications in hydrology, *Hydrol. Sci. J.*, doi:10.1080/02626667.2016.1162953.
- Krawczyk, M., and E. Reuben (2012), (Un) Available upon request: Field experiment on researchers' willingness to share supplementary materials, *Account. Res.*, *19*(3), 175–186, doi:10.1080/08989621.2012.678688.
- Kumar, M. (2009), Toward a hydrologic modeling system, PhD dissertation, 251 pp., The Pa. State Univ.
- Kumar, M., and C. Duffy (2015), Exploring the role of domain partitioning on efficiency of parallel distributed hydrologic model simulations, *J. Hydrogeol. Hydrol. Eng.*, *4*(1), 1–12, doi:10.4172/2325-9647.1000119.
- Kumar, M., D. Marks, J. Dozier, M. Reba, and A. Winstral (2013), Evaluation of distributed hydrologic impacts of temperature-index and energy-based snow models, *Adv. Water Resour.*, *56*, 77–89, doi:10.1016/j.advwatres.2013.03.006.
- Leonard, L., and C. J. Duffy (2013), Essential terrestrial variable data workflows for distributed water resources modeling, *Environ. Modell. Software*, *50*, 85–96, doi:10.1016/j.envsoft.2013.09.003.
- Leonard, L., and C. J. Duffy (2014), Automating data-model workflows at a level 12 HUC scale: Watershed modeling in a distributed computing environment, *Environ. Modell. Software*, *61*, 174–190, doi:10.1016/j.envsoft.2014.07.015.
- Leonard, L., and C. J. Duffy (2016), Visualization workflows for level-12 HUC scales: Towards an expert system for watershed analysis in a distributed computing environment, *Environ. Modell. Software*, *78*, 163–178, doi:10.1016/j.envsoft.2016.01.001.
- Lichtner, P. C., G. E. Hammond, C. Lu, S. Karra, G. Bisht, B. Andre, R. T. Mills, and J. Kumar (2013), PFLOTRAN. [Available at <http://www.pflotran.org>.]
- Lowry, C. S., and M. N. Fienen (2013), CrowdHydrology: Crowdsourcing hydrologic data and engaging citizen scientists, *Ground Water*, *51*(1), 151–156, doi:10.1111/j.1745-6584.2012.00956.x.
- Maxwell, R. M., and S. J. Kollet (2008), Interdependence of groundwater dynamics and land-energy feedbacks under climate change, *Nat. Geosci.*, *1*(10), 665–669, doi:10.1038/ngeo315.
- Maxwell, R. M., M. Putti, S. Meyerhoff, J.-O. Delfs, I. M. Ferguson, V. Ivanov, J. Kim, O. Kolditz, S. J. Kollet, and M. Kumar (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.
- Mayernik, M. S., S. Callaghan, R. Leigh, J. Tedds, and S. Worley (2015), Peer review of datasets: When, why, and how, *Bull. Am. Meteorol. Soc.*, *96*(2), 191–201, doi:10.1175/BAMS-D-13-00083.1.
- McNutt, M., K. Lehnert, B. Hanson, B. A. Nosek, A. M. Ellison, and J. L. King (2016), Liberating field science samples and data, *Science*, *351*(6277), 1024–1026, doi:10.1126/science.aad7048.
- Mirus, B. B., K. Loague, N. C. Cristea, S. J. Burges, and S. K. Kampf (2011), A synthetic hydrologic response dataset, *Hydrol. Processes*, *25*(23), 3688–3692, doi:10.1002/hyp.8185.
- Moreau, L. (2010), The foundations for provenance on the web, *Found. Trends Web Sci.*, *2*(2/3), 99–241, doi:10.1561/1800000010.
- Nekrutenko, A., and J. Taylor (2012), Next-generation sequencing data interpretation: Enhancing reproducibility and accessibility, *Nat. Rev. Genet.*, *13*(9), 667–672, doi:10.1038/nrg3305.
- Niu, J., C. Shen, S.-G. Li, and M. S. Phanikumar (2014), Quantifying storage changes in regional Great Lakes watersheds using a coupled subsurface land surface process model and GRACE, MODIS products, *Water Resour. Res.*, *50*, 7359–7377, doi:10.1002/2014WR015589.
- Noël, P., A. N. Rousseau, C. Paniconi, and D. F. Nadeau (2014), An algorithm for delineating and extracting hillslopes and hillslope width functions from gridded elevation data, *J. Hydrol. Eng.*, *19*(2), 366–374, doi:10.1061/(ASCE)HE.1943-5584.0000783.
- Nosek, B. A., et al. (2015), Promoting an open research culture, *Science*, *348*(6242), 1422–1425, doi:10.1126/science.aab2374.
- National Science Foundation (2015), Public Access Plan: Today's data, tomorrow's discoveries. [Available at <https://www.nsf.gov/pubs/2015/nsf15052/nsf15052.pdf>.]
- Paola, C., E. Fofoula-Georgiou, W. E. Dietrich, M. Hondzo, D. Mohrig, G. Parker, M. E. Power, I. Rodriguez-Iturbe, V. Voller, and P. Wilcock (2006), Toward a unified science of the Earth's surface: Opportunities for synthesis among hydrology, geomorphology, geochemistry, and ecology, *Water Resour. Res.*, *42*, W03S10, doi:10.1029/2005WR004336.
- Peckham, S. D. (2014), The CSDMS Standard Names: Cross-domain naming conventions for describing process models, data sets and their associated variables, paper presented at International Environmental Modelling Software Society (IEMSs), San Diego, Calif.
- Penn, C. A., L. A. Bearup, R. M. Maxwell, and D. W. Clow (2016), Numerical experiments to explain multiscale hydrological responses to mountain pine beetle tree mortality in a headwater watershed, *Water Resour. Res.*, doi:10.1002/2015WR018300, in press.
- Persson, P.-O., and G. Strang (2004), A simple mesh generator in MATLAB, *SIAM Rev.*, *46*, 329–345.

- PIHM (2015), Pennsylvania State Integrated Hydrologic Model version 2.2, *Zenodo*, doi:10.5281/zenodo.31438.
- Qu, Y. (2005), An integrated hydrologic model for multi-process simulation using semi-discrete finite volume approach, PhD thesis, 136 pp., Pa. State Univ.
- Qu, Y., and C. J. Duffy (2007), A semidiscrete finite volume formulation for multiprocess watershed simulation, *Water Resour. Res.*, *43*, W08419, doi:10.1029/2006WR005752.
- Roulstone, I., and J. Norbury (2013), Climate-model misgivings, Project Syndicate, 2013.
- Rousseau, A. N., J.-P. Fortin, R. Turcotte, A. Royer, S. Savary, F. Quévy, P. Noël, and C. Paniconi (2011), PHYSITEL: A specialized GIS for supporting the implementation of distributed hydrological models, *Water News - Can. Water Resour. Assoc.*, *31*(1), 18–20.
- Rousseau, A. N., A. Royer, and S. Tremblay (2013), Hydroinformatics of watershed hydrology at INRS—Development framework and examples of data processing, editing, display, and transfer algorithms implemented in our software, paper presented at Open Geospatial Consortium, Hydrology Domain Working Group, INRS-ETE, Québec, Que., 20 June.
- Rousseau, A. N., A. P. Alvarez, X. Yu, S. Savary, and C. Duffy (2015), Comparison of two conceptually different physically-based hydrological models—Looking beyond streamflows, Abstract H23E-1624 presented at 2015 Fall Meeting, AGU, San Francisco, Calif., 14–18 Dec.
- Shepardson, D. P., B. Wee, M. Priddy, L. Schellenberger, and J. Harbor (2007), What is a watershed? Implications of student conceptions for environmental science education and the national science education standards, *Sci. Educ.*, *91*(4), 554–578, doi:10.1002/sce.20206.
- Shi, Y., K. J. Davis, C. J. Duffy, and X. Yu (2013), Development of a coupled land surface hydrologic model and evaluation at a critical zone observatory, *J. Hydrometeorol.*, *14*, 1401–1420, doi:10.1175/JHM-D-12-0145.1.
- Shi, Y., K. J. Davis, F. Zhang, C. J. Duffy, and X. Yu (2014), Parameter estimation of a physically based land surface hydrologic model using the ensemble Kalman filter: A synthetic experiment, *Water Resour. Res.*, *50*, 706–724, doi:10.1002/2013WR014070.
- Shi, Y., K. J. Davis, F. Zhang, C. J. Duffy, and X. Yu (2015a), Parameter estimation of a physically-based land surface hydrologic model using an ensemble Kalman filter: A multivariate real-data experiment, *Adv. Water Resour.*, *83*, 421–427, doi:10.1016/j.advwatres.2015.06.009.
- Shi, Y., D. C. Baldwin, K. J. Davis, X. Yu, C. J. Duffy, and H. Lin (2015b), Simulating high-resolution soil moisture patterns in the Shale Hills Watershed using a land surface hydrologic model, *Hydrol. Processes*, *29*(21), 4624–4637, doi:10.1002/hyp.10593.
- Skaggs, T. H., M. H. Young, and J. A. Vrugt (2015), Reproducible research in vadose zone sciences, *Vadose Zone J.*, *14*(10), doi:10.2136/vzj2015.06.0088.
- Stewart, I. (2016), Sustainable geoscience, *Nat. Geosci.*, *9*, 262, doi:10.1038/ngeo2678.
- Stodden, V. C. (2011), Trust your science? Open your data and code, *Amstat News*, *409*, 21–22.
- Sutanudjaja, E., et al. (2015), The HyperHydro (H²) experiment for comparing different large-scale models at various resolutions, Abstract presented at 2015 Fall Meeting, AGU.
- Tarboton, D. G., et al. (2014), HydroShare: Advancing collaboration through hydrologic data and model sharing, in *Proceedings of the 7th International Congress on Environmental Modelling and Software*, edited by D. P. Ames, N. W. T. Quinn, and A. E. Rizzoli, Int. Environ. Modell. and Software Soc. (iEMSS), San Diego, Calif. [Available at http://www.iemss.org/sites/iemss2014/papers/iemss2014_submission_243.pdf].
- Turcotte, R., J.-P. Fortin, A. N. Rousseau, S. Massicotte, and J.-P. Villeneuve (2001), Determination of the drainage structure of a watershed using a digital elevation model and a digital river and lake network, *J. Hydrol.*, *240*, 225–242, doi:10.1016/S0022-1694(00)00342-5.
- Turcotte, R., A. N. Rousseau, J.-P. Fortin, and J.-P. Villeneuve (2003), A process-oriented, multiple-objective calibration strategy accounting for model structure, in *Calibration of Watershed Models*, *Water Sci. & Appl. Ser.*, vol. 6, edited by Q. Duan, et al., pp. 153–163, AGU, Washington, D. C., doi:10.1002/9781118665671.ch11.
- Turcotte, R., L.-G. Fortin, V. Fortin, J.-P. Fortin, and J.-P. Villeneuve (2007), Operational analysis of the spatial distribution and the temporal evolution of the snowpack water equivalent in southern Québec, Canada, *Hydrol. Res.*, *38*(3), 211–234, doi:10.2166/nh.2007.009.
- U.S. Geological Survey (2015), Data Management Policy References. [Available at <http://www.usgs.gov/datamanagement/policyreferences.php>, Accessed on 30 Dec. 2015.]
- Wang, R., M. Kumar, and D. Marks (2013), Anomalous trend in soil evaporation in a semi-arid, snow-dominated watershed, *Adv. Water Resour.*, *57*, 32–40, doi:10.1016/j.advwatres.2013.03.004.
- Yang, J., T. Graf, M. Herold, and T. Ptak (2013), Modelling the effects of tides and storm surges on coastal aquifers using a coupled surface–subsurface approach, *J. Contam. Hydrol.*, *149*, 61–75, doi:10.1016/j.jconhyd.2013.03.002.
- Yu, X. (2015a), PIHM input to benchmarking V-catchment, doi:10.6084/m9.figshare.1328521.
- Yu, X. (2015b), PIHM input of Shale Hills Watershed, doi:10.6084/m9.figshare.1506789.
- Yu, X. (2016a), Observed streamflow of Shale Hills Watershed in 1974, doi:10.6084/m9.figshare.2069037.
- Yu, X. (2016b), Plot PIHM mesh file in 3D, doi:10.6084/m9.figshare.2069031.
- Yu, X. (2016c), Plot the modeled streamflow of V-shaped Catchment, doi:10.6084/m9.figshare.2069040.
- Yu, X. (2016d), Plot the observed and modeled streamflow of Shale Hills, doi:10.6084/m9.figshare.2069034.
- Yu, X., G. Bhatt, C. Duffy, and Y. Shi (2013), Parameterization for distributed watershed modeling using national data and evolutionary algorithm, *Comput. Geosci.*, *58*, 80–90, doi:10.1016/j.cageo.2013.04.025.
- Yu, X., G. Bhatt, C. J. Duffy, D. H. Wardrop, R. G. Najjar, A. C. Ross, and M. Rydzik (2015a), A coupled surface–subsurface modeling framework to assess the impact of climate change on freshwater wetlands, *Clim. Res.*, *66*(3), 211–228, doi:10.3354/cr01348.
- Yu, X., C. Duffy, Y. Gil, L. Leonard, G. Bhatt, and E. Thomas (2015b), Cyber-innovated watershed research at the Shale Hills Critical Zone Observatory, *Syst. J. IEEE, PP*(99), 1–12, doi:10.1109/JSYST.2015.2484219.
- Yu, X., A. Lamačová, C. J. Duffy, P. Krám, J. Hruška, T. White, and G. Bhatt (2015c), Modeling the long-term water yield effects of forest management in a Norway spruces forest, *Hydrol. Sci. J.*, doi:10.1080/02626667.2014.897406.
- Zhang, Y., R. Slingerland, and C. J. Duffy (2016), Fully-coupled hydrologic processes for modeling landscape evolution, *Environ. Model. Software*, doi:10.1016/j.envsoft.2016.04.014.