Revealing causal controls of storage-streamflow relationships with a data-centric Bayesian framework combining machine learning and process-based modeling

Wen-Ping Tsai[1], Kuai Fang[2], Xinye Ji[1], Kathryn Lawson[1], and Chaopeng Shen[1]

[1]Pennsylvania State University
[2]Stanford University

November 26, 2022

Abstract

Some machine learning (ML) methods such as classification trees are useful tools to generate hypotheses about how hydrologic systems function. However, data limitations dictate that ML alone often cannot differentiate between causal and associative relationships. For example, previous ML analysis suggested that soil thickness is the key physiographic factor determining the storage-streamflow correlations in the eastern US. This conclusion is not robust, especially if data are perturbed, and there were alternative, competing explanations including soil texture and terrain slope. However, typical causal analysis based on process-based models (PBMs) is inefficient and susceptible to human bias. Here we demonstrate a more efficient and objective analysis procedure where ML is first applied to generate data-consistent hypotheses, and then a PBM is invoked to verify these hypotheses. We employed a surface-subsurface processes model and conducted perturbation experiments to implement these competing hypotheses and assess the impacts of the changes. The experimental results strongly support the soil thickness hypothesis as opposed to the terrain slope and soil texture ones, which are co-varying and coincidental factors. Thicker soil permits larger saturation excess and longer system memory that carries wet season water storage to influence dry season baseflows. We further suggest this analysis could be formalized into a novel, data-centric Bayesian framework. This study demonstrates that PBM present indispensable value for problems that ML cannot solve alone, and is meant to encourage more synergies between ML and PBM in the future.

Revealing causal controls of storage-streamflow relationships with a data-centric Bayesian framework combining machine learning and process-based modeling

Wen-Ping Tsai[1], Kuai Fang[1,2], Xinye Ji[1,3], Kathryn Lawson[1], Chaopeng Shen[1,*]

[1] Civil and Environmental Engineering, Pennsylvania State University, University Park, Pennsylvania, USA
[2] Earth System Science, Stanford University, USA
[3] Shenzhen State High-Tech Industrial Innovation Center, Shenzhen, China, 518000

Abstract

Some machine learning (ML) methods such as classification trees are useful tools to generate hypotheses about how hydrologic systems function. However, data limitations dictate that ML alone often cannot differentiate between causal and associative relationships. For example, previous ML analysis suggested that soil thickness is the key physiographic factor determining the storage-streamflow correlations in the eastern US. This conclusion is not robust, especially if data are perturbed, and there were alternative, competing explanations including soil texture and terrain slope. However, typical causal analysis based on process-based models (PBMs) is inefficient and susceptible to human bias. Here we demonstrate a more efficient and objective analysis procedure where ML is first applied to generate data-consistent hypotheses, and then a PBM is invoked to verify these hypotheses. We employed a surface-subsurface processes model and conducted perturbation experiments to implement these competing hypotheses and assess the impacts of the changes. The experimental results strongly support the soil thickness hypothesis as opposed to the terrain slope and soil texture ones, which are co-varying and coincidental factors. Thicker soil permits larger saturation excess and longer system memory that carries wet season water storage to influence dry season baseflows. We further suggest this analysis could be formalized into a novel, data-centric Bayesian framework. This study demonstrates that PBM present indispensable value for problems that ML cannot solve alone, and is meant to encourage more synergies between ML and PBM in the future.

Keywords:

Machine learning (ML), process-based model (PBM), streamflow-storage relationships, data-centric

[*] Corresponding author: cshen@engr.psu.edu

1. Background

Basin water storage has deep connections with streamflow (Reager et al., 2014; Fang and Shen, 2017). Hence terrestrial water storage anomalies (TWSA) data could, under certain circumstances, be used to increase flood forecast lead time (Reager et al., 2015). From a physical hydrologic point of view, more water stored in a basin could mean a higher groundwater table or wetter soils which lead to more runoff source areas (Dingman, 2015). The storage-streamflow relationship is also important for predicting baseflow (Thomas et al., 2013) and related ecosystem (Poff and Allan, 1995) and water supply issues. The issue is that these relationships vary widely in space. Fang and Shen (2017) (hereafter named FS17, more description in Section 2) conducted an analysis of the correlation between TWSA annual extrema and different streamflow percentiles in a year, and found very interesting patterns of these correlations over the continental United States (CONUS). The correlations between TWSA annual extrema and high-percentile flows are strong in certain parts of the CONUS, e.g., the southeastern coastal plains and northern great plains, but are weak in other areas such as the Appalachian Plateau, northern Indiana, and Florida. *Why are there wildly different storage-streamflow relationships, i.e., what physical factors caused them?* Our limited understanding of this question hampered our use of water storage and groundwater data in flood forecasting.

In general, to answer "*why*" questions such as the one raised above, one could resort to two avenues: process-based models (PBMs) or data-driven analysis. They are often regarded as two separate roads that do not cross. PBMs embody our *beliefs* about how the system functions. We can use PBMs to conduct numerical experiments to assess causal relationships, as we can alter measurable physical factors to directly examine their impacts on the outputs. We typically employ a 'model-centric' framework, where we (i) deploy some prior distributions or beliefs of model structures; (ii) create an ensemble of model simulations (with different parameter sets, inputs, or model structures); (iii) confront these models with observations by evaluating likelihood functions either formally or informally by visually examining the outcomes; and (iv) identify the model(s) that best describe(s) the data. It is easy to see that paradigms like model calibration (Vrugt et al., 2003) or Monte Carlo Markov Chain (Vrugt et al., 2009) fit into this framework. Moreover, numerical experiments where the modelers perturb model physics on an ad-hoc basis (e.g., (Maxwell and Condon, 2016; Shen et al., 2016; Ji et al., 2019)) could also be placed in this framework. Potential issues with this framework are that it can be both subjective and inefficient, as many competing hypotheses remain un-tested. The priors are often based on one's own beliefs, and one needs to throw a huge amount of simulations to capture the plausible model structure. It has been argued that hydrologic models are necessarily degenerate (Nearing et al., 2016) and even sampling exhaustively from its parameter distribution does not capture the whole possible model space.

In contrast to PBMs, various interpretable data mining approaches could be used to generate possible explanations, or "hypothesis" in machine learning language (Russell and Norvig, 2009), of an observed behavior. For example, the weights from linear regression could inform us of the relative importance of factors. Classification and regression tree (CART) (Breiman et al., 1984; Mitchell, 1997), which iteratively separates data points based on predictors and their thresholds, is another explanatory tool that has often been employed. For example, Verhougstraete et al., (2015) used the first level split in a CART model to draw the conclusion that septic systems are the primary driver of fecal bacteria levels in 64 US rivers. An advantage of data mining approaches is that they are highly efficient to execute compared to PBMs, and the models they generate are already consistent with data. They also carry the appeal of relying less on subjective assumptions and model choices.

However, the “Achilles heel” for data mining as an explanatory tool is arguably their inability to distinguish between causal and associated relationships. If we had a large enough training dataset that covered all possible combinations of physical factors, data mining should theoretically be able to extract the causal factor. However, we are limited by the combinations that exist in the real world and for which we have data, posing limits on the power of data. Naturally, one might wonder if PBMs’ strength in causality analysis could be exploited to complement data mining algorithms.

Recently, there have emerged increasing interest in combining physics with data-driven models. One could adopt a variety of methods loosely termed “physics-guided machine learning” (PGML) or “theory-guided machine learning” (Ganguly et al., 2014; Karpatne et al., 2017; Jia et al., 2019; Yang et al., 2019), such as modifying the loss function to accommodate physical constraints (Jia et al., 2019) or pre-training a ML model using PBM outputs (Jia et al., 2018). These constructive ideas have made ML more robust and have enriched our means of investigations. Nevertheless, PGML frameworks have not taken advantage of PBM’s ability to conduct experiments and assess causes and effects.

Here we propose that the evaluation of competing hypotheses could be accomplished by running numerical experiments with a PBM to utilize the physics encoded in the PBM (Figure 2), as an example of the alternative research avenue proposed earlier (Shen et al., 2018). We then compare the probability of each hypothesis and reject those with low probability. Because this framework first starts with data, we call it a data-centric framework, in contrast to a conventional mode-centric Bayesian framework where a model’s inputs and parameters are perturbed and the posterior probability of each realization is calculated. We will use the storage-streamflow question to showcase the effectiveness of this framework and help us understand the main controlling factors of streamflow in these regions to inspire best modeling practices. This work is a first exploration of this particular method of coupling data-driven hypothesis with process-based modeling capabilities, and by no means do we indicate this method is optimal or the most efficient.

In the following, we first provide some background for the case study of streamflow-storage correlations and the competing hypotheses that explain them (Section 2). Then we describe the process-based model and the experimental setup (Section 3.2 and Section 3.4). We make sure the model produces reasonable hydrologic dynamics (Section 4.2), and then finally we use the perturbation experiments to test the competing hypotheses from ML (Section 4.3).

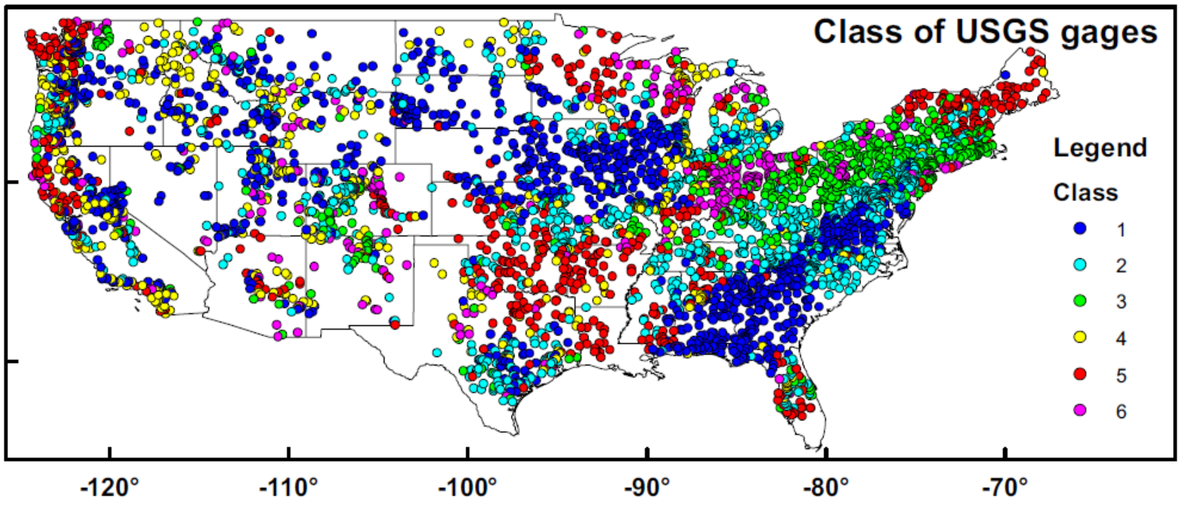
2. The background story

2.1 The Storage-Streamflow-Correlation Spectrum

In FS17 we introduced a hydrologic signature termed the Storage-Streamflow-Correlation Spectrum (SSCS), which quantifies how water storage is correlated with streamflow at different flow regimes. Concisely, SSCS is the collection of Pearson's correlation coefficients (R) between annual extrema (peaks or troughs) of the terrestrial water storage anomalies (TWSA) and different streamflow percentiles (15 percentiles extracted are: {0.5%, 1%, 2%, 5%, 10%, 20%, 50%, 60%, 70%, 80%, 90%, 95%, 98%, 99%, 99.5%}) in a window around the extrema for the same basin. The correlations are calculated on an annual scale, using the water year (the 12-month period from October 1 through September 30 of the following year). The study period of FS17 is from 1 October 2002 to 31 September 2012. Treating each flow percentile as a "band", we obtained a correlation "spectrum". The SSCS gives a snapshot of the correlations across all bands, as compared to previous studies that focused only on high flow regimes.

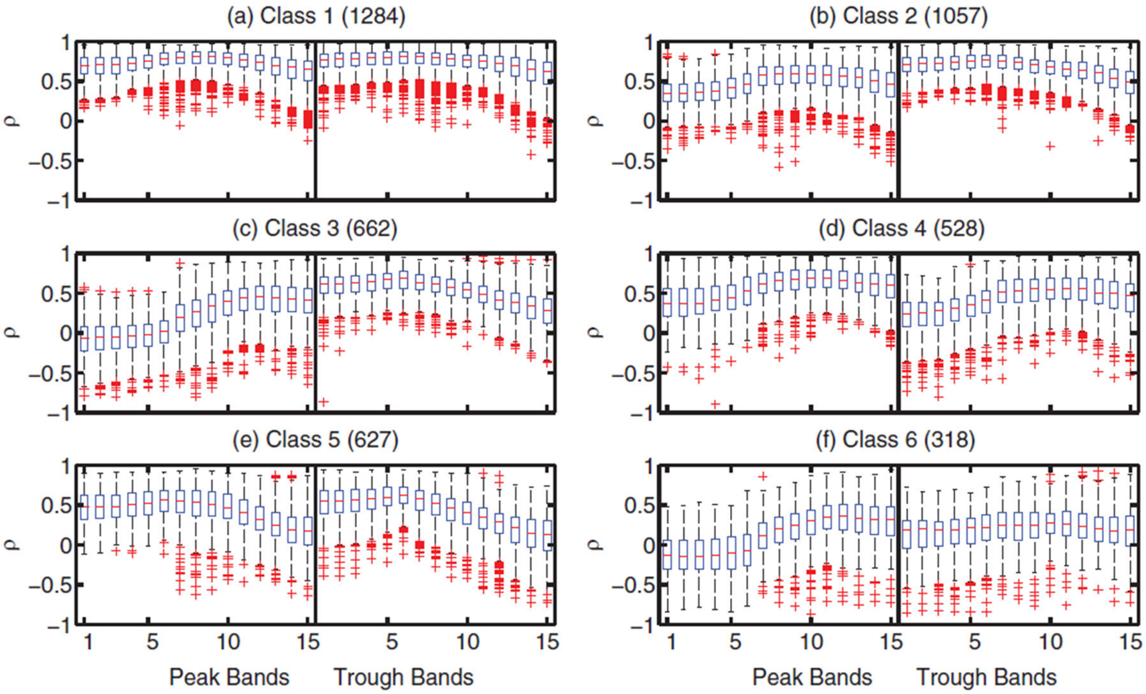
If streamflow is disconnected from storage, e.g., when most rainfall runs off or evaporates directly without entering the subsurface, the system would exhibit low correlation between flows and storage during peak flows. Generally, the high-flow bands have lower R than low-flow bands because peak streamflows result from large storms whose magnitudes are poorly correlated to water storage. In contrast, if groundwater exerts a significant influence over streamflow, we expect the correlation to be higher. A high correlation between TWSA peaks and low flows indicates a long system memory: when such basins receive plenty of precipitation in the wet season, the excess storage is carried over the seasons and is reflected in low flows. Therefore, SSCS gives us a window of observation into how varied surface and subsurface hydrologic systems function. Please see FS17 for more details.

When applying the SSCS over the continental United States (CONUS), a large variety of SSCS behaviors emerged (FS17). To facilitate our interpretation, we clustered these responses into 6 different classes using K-means and a distance measure (the Euclidean distance in the SSCS space). The correlation values for different classes and the spatial distribution of classes are shown in Figure 1. We can clearly observe regional clusters and spatial gradients in the SSCS patterns. Class #1 was described as ‘‘full-spectrum responsive’’ since it had the highest correlations and the smallest variability across all SSCS bands. Class #1 concentrated on the southeast coastal plains and northern great plains. Class #2 and #3 catchments had weaker SSCS values and were concentrated along the northern Appalachian Plateau. For Class #3, in peak-TWSA bands, streamflow-storage correlation was low for flow percentiles below 20%, but higher for percentiles above 60%; in trough-TWSA bands, there were high streamflow-storage correlations at percentiles below 60%, but correlations were a little lower for high streamflow percentiles (80% above). Class #2 can be considered a transition type between class #1 and class #3.



Class of USGS gages
Legend
Class
● 1
● 2
● 3
● 4
● 5
● 6
40°
30°
-120°
-110°
-100°
-90°
-80°
-70°

(A)



(a) Class 1 (1284)
(b) Class 2 (1057)
(c) Class 3 (662)
(d) Class 4 (528)
(e) Class 5 (627)
(f) Class 6 (318)
ρ
1
0.5
0
−0.5
−1
1
5
10
15
Peak Bands
Trough Bands

(B)

Figure 1. Class map (A) and boxplots of the SSCS for Class #1 to Class #6 (B). The boxes contain 25–75% percentiles, and the crosses are those considered outliers (Reprinted from FS17 with permission).

2.2 Explaining the controls of SSCS

When observing the large spatial gradients of SSCS classes over CONUS in Figure 1, one cannot help asking, "***what causes the SSCS behavior to differ between Appalachia and the coastal plains?***", which was the central question of this study. FS17 employed CART to learn simple and interpretable decision rules (the split criteria and thresholds) from the data. Focusing on the contrasts between the Appalachian basins and the basins on the southeast coastal plains, FS17 trained a specific CART model to predict the distances of basins to class centers in SSCS space. They used a number of predictors including the aridity index, depth to bedrock, rainfall seasonality, and the fraction of precipitation as snow (supporting information Table S1 in FS17). In other words, they asked what factors made the two clusters of basins different in terms of their SSCS patterns. From this ad hoc tree, CART automatically identified soil thickness (RockDep), obtained by merging soils-survey-based depth to bedrock with bedrock depth simulated by a geomorphological model (Pelletier et al., 2016)) as the main difference between the two types of streamflow-storage correlation patterns.

The problem of learning an optimal CART is that CART is not robust. This can be mitigated by training multiple trees in an ensemble as in random forest (RF) (Ho, 1995), where the features and samples are randomly sampled with replacement. The RF generalizes from the CART and provides an estimation of probability. While RF models are more robust and can be used to infer probabilities, they are more difficult for humans to interpret.

While the RockDep explanation does make physical sense, it could be dangerous to take this hypothesis as the truth. First, even though soil thickness appeared to be the stronger explanatory model, there could be other slightly weaker but nonetheless valid models. We have yet to explore what would happen if we slightly alter the training dataset. Because we rely on available data, the results may be dependent on a few data points that critically cover certain parts of the input space. However, such critical data points may happen to be missing in our training data; the robustness of the model has not been established.

3. Methods and Datasets

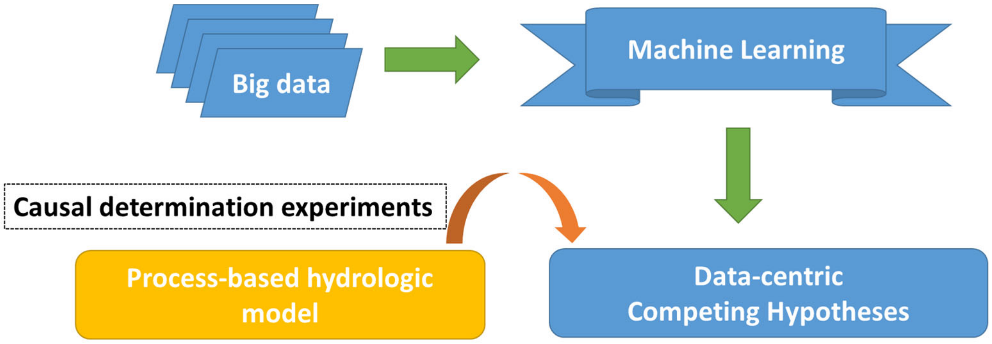

Big data
Machine Learning
Causal determination experiments
Process-based hydrologic model
Data-centric Competing Hypotheses

Figure 2. A framework of integrating PBM and machine learning

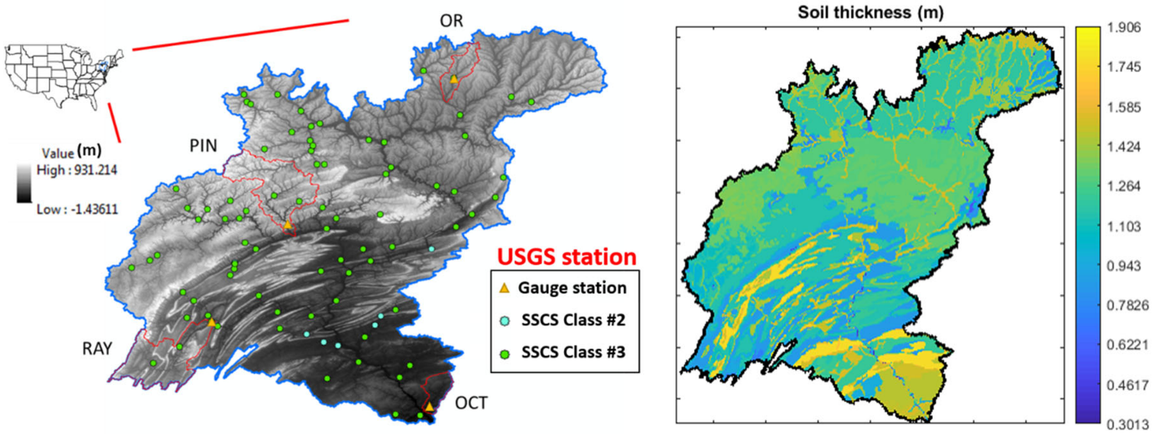
To answer our central question, here we propose a novel framework that combines the strengths of machine learning and process-based modeling. In this framework, machine learning first presents competing hypotheses and assign them prior probabilities. Then, we construct numerical perturbation experiments with a process-based model to implement and test the hypotheses (Figure 2). The testing of the hypotheses could be achieved by visual examination of the outcome of the experiments, or via a more quantitative Bayesian approach.

3.1 Study area - Susquehanna River Basin (SRB)

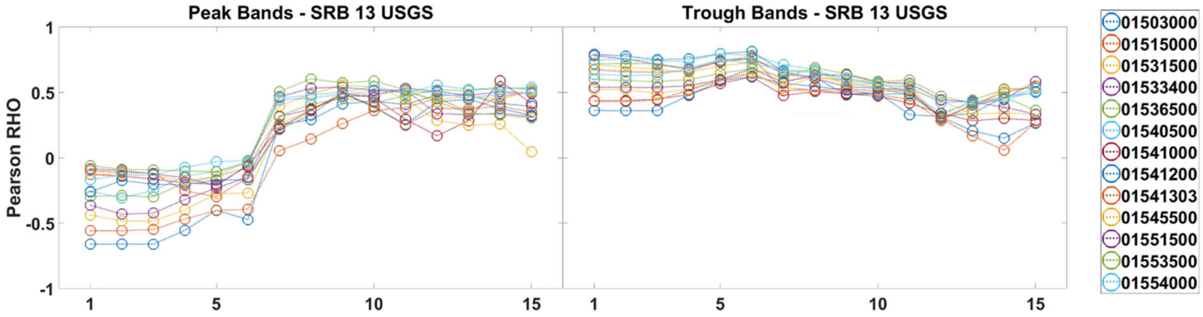
The Susquehanna River (watershed area: 71,225 km^2) is a major river located in the northeastern and mid-Atlantic United States (Figure 3a), which has historically been the source of many instances of flooding damagealong the main river floodplains (Yarnal et al., 1997; May, 2011).

The basin spans the physiographic provinces of the Appalachian Plateau, Piedmont, Valley and Ridge, and coastal plains. In general, most of the northern subbasins of the SRB consist of mountains mantled by thin soils which are mostly thinner than 2 meters (Figure 3b). We show the SSCS behaviors of 13 randomly selected subbasins in the Susquehanna River Basin. We found that the all 13 stations in the Susquehanna River Basin belong to either Class #2 or Class #3 (Figure 3c, the original pattern of SSCS from 13 USGS gauge stations is similar to class #3).

We further chose 4 subbasins (Figure 4a), namely, the Otselic River basin (OR), the Pine Creek basin (PIN), the Raystown Branch Juniata River basin (RAY), and the Octoraro Creek basin (OCT) in the south to create process-based hydrologic models. Both soils survey data and global modeled soil thickness data were used to parameterize soil thickness: in most of the basin where the bedrock is within the limit of the soils survey depth (1.52 m), the RockDep attribute in SSURGO (NRCS, 2010) was used; outside of these areas, we used the average soil and sedimentary layer thickness from Pelletier et al., (2016), which has global coverage with 1 km resolution. Among the subbasins modeled, OR and PIN are headwater basins in the Appalachian Plateau, RAY is a headwater subbasin in the Valley and Ridge physiographic division, and OCT is near the coastal plains. OCT has a visibly larger soil thickness.



(a) Study area – Susquehanna River Basin. ***(b) Soil thickness***



(c) SSCS in Susquehanna River Basin for 13 USGS gauge station (station number is shown in legend).

Figure 3. Study area, Susquehanna River Basin (SRB). Main class of SRB observation data is class #3 and #2 in FS17.

3.2 Process-based hydrologic model

To be able to conduct causal experiments, we employed the Process-based Adaptive Watershed Simulator coupled with the Community Land Model (PAWS+CLM) (Shen and Phanikumar, 2010; Shen et al., 2013, 2014, 2016, Ji et al., 2015, 2019; Niu et al., 2017; Ji and Shen, 2018; Fang et al., 2019). First introduced in Shen and Phanikumar (2010), the model was coupled to the Community Land Model (CLM) (Collins et al., 2006; Dickinson et al., 2006; Oleson et al., 2010; Lawrence et al., 2011) which describes the land surface and vegetation dynamics (Shen et al., 2013). The PAWS model has been used to explain the relative importance of different controlling processes on hydrologic and ecosystem dynamics. CLM incorporates comprehensive physical and biogeochemical processes including vapor and momentum transfer, surface radiative transfer, soil heat transfer, freeze-thaw phase changes, and biochemical photosynthesis, as well as plant carbon and nitrogen cycles (Shen et al., 2014). PAWS+CLM inherits the land surface processes from CLM, including surface energy fluxes, ET, vegetation growth, and carbon cycling, while solving physically-based conservative laws for flow processes including 2D overland flow, quasi-3D subsurface (soil and groundwater) flow, vectorized channel network, and the exchanges among these domains. The flow module starts with throughfall, stemflow, and snowmelt as the precipitation inputs, and converts the CLM-computed evapotranspiration term into a sink. The surface water layer is divided into the flow domain, which can flow laterally, and the ponding domain, which exchanges with the main soil column and does not circulate laterally. The flow domain water is routed downstream as overland flow, described by 2D diffusive wave equation (DWE). Infiltrated water is governed by the Richards equation. Water reaching the phreatic water table may move laterally, as described by Dupuit-Forchheimer flow in an unconfined aquifer. 1D columns of vertical soil flow are coupled to the saturated lateral flow at the bottom. The confined aquifers below are described by a 3D saturated groundwater flow equation. The channel flow is governed by DWE in a 1D cascade network. More information about PAWS can be found in Shen et al. (2016).

3.3 Configuration of the hydrologic model

In this study, a 1040 m x 1040 m horizontal grid was used to discretize the domain. Precipitation and climate forcing data used in PAWS+CLM were obtained from the North American Land Data Assimilation System (NLDAS) (Mitchell, 2004). Information from the Soil Survey Geographic Database (SSURGO) was used to provide initial values for the soil properties. In PAWS+CLM, we extracted topographic information from the National Elevation Dataset (30 m) to parameterize the river bed elevations, and used the mean elevation to parameterize the gridcell elevation (Shen et al., 2016). The climatic forcing datasets that comefrom NLDAS are on an hourly basis.

The channel network is represented by an explicit, vectorized channel network for larger rivers and the implicit, gridded overland flow for smaller headwater streams. As an advance of PAWS+CLM, the channel network topology is now established based on the National

Hydrography Dataset Plus Version 2 (NHDPlus V2) shapefiles. In NHDPlus V2, each segment is encoded with a unique ID number and the downstream ID. Combing through this connectivity information, our pre-processing package traces the rivers from downstream to upstream and records the river distances of each segment. The available channels from NHD are vastly greater than what can be explicitly represented in the vectorized channel network in the model. In previous work, the selection of the explicitly modeled streams was manual. We have now implemented an automatic selection procedure: our pre-processing utility iteratively selects the longest rivers from the candidate pool built from NHDPlus V2, so that the total selected river length satisfies a prescribed river density (river length : basin area). Based on these explicitly represented rivers, we then establish a network structure, recording names of the streams, network topology, upstream/downstream nodes in the hierarchy, boundary condition types (headwater, inflow, connecting streams, or outflow), tributaries, and locations of confluences. For each explicitly modeled river, the discretization procedure evenly distributes the river polyline into river cells. We then overlay the river cell with high resolution DEM and groundwater data, extracting information, e.g. bank and bed elevation (inferred through regional regression equation), during discretization (Shen et al., 2016).

In PAWS the soil water retention and unsaturated hydraulic conductivity are parameterized using the van Genuchten formulation. To obtain spatially distributed van Genuchten parameters, we incorporated a range of well-established pedotransfer functions (PTFs) (Guber et al., 2009), and the Rosetta (Schaap et al., 2001) program which employs a hierarchy of PTFs, ranging in complexity from a soil textural lookup table to algorithms based on Artificial Neural Networks (ANN). We also exported soil textural information (sand, clay, and silt percentages), bulk density, and water contents from soil horizon data from the SSURGO database (NRCS, 2010) into Rosetta, wherever they were available. Rosetta was then used to predict van Genuchten parameters, and the results were subsequently read into PAWS. Normally, we chose the ‘best possible model’ option in Rosetta. The SSURGO database contains fine resolution (1:24,000 map scale) soil type maps, which are encoded as ‘map unit’ keys (mukey). A mukey value serves as an index key to the SSURGO relational databases that detail the characteristics of that soil type. A mukey may contain several ‘soil components’, each taking up a certain fraction of the map unit. Every component then describes the vertical soil horizons and their depths.

Even with the help of the pedotransfer functions, process-based hydrologic model parameters need to be further adjusted or calibrated. The SRB is large, and it is difficult to perform calibration for the whole basin. We thus defined our objective function as the mean of the Nash-Sutcliffe model efficiency coefficients (Nash and Sutcliffe, 1970) for the four subbasins. This way, the resulting parameter set may not produce the best achievable performance for each subbasin, but presents a balance between them for the whole basin. Model performance was evaluated against USGS streamflow records.

3.4 Competing hypotheses and the implementation of perturbation experiments

To identify potential competing hypotheses, we first ran both CART analysis for southeastern and Appalachian basins with multiple random seeds and randomized removal of training data points (basins). Then we further ran RF analysis with an expanded list of attributes. With CART, we considered all basin physiographic parameters that were deemed as important for SSCS in FS17, including: RockDep, sand, slope, soil bulk density, watershed percent agriculture, watershed percent developed, and standard deviation of elevation. In FS17, we employed sand and clay as representatives for soil texture and removed silt, since they add up to one. In the present analysis we also followed this practice. We then implemented changes in these factors via perturbing corresponding parameters in the process-based model. Essentially, we first replaced the values of these factors in the SRB by their counterparts from the Southeastern CONUS, and ran experiments to determine their individual impacts on the SSCS classes. We also considered the combinatory impacts of these factors by altering them at the same time.

Some climatic variables such as relative humidity, annual precipitation, and fraction of precipitation as snow could overtake as the top-level split, but are ignored in the manual CART analysis because we are interested in the relative impacts of physical basin parameters. We nonetheless included them in the RF model and PBM perturbation experiments by replacing forcing data on the SRB with those from some locations on the coastal plains, to compare their impacts with the physical basin parameters.

One of the important physical basin parameters is soil thickness. The difference in average soil thickness between then thinly-mantled Appalachian basins and their southeastern neighbors is about 30 meters. Hence, for the perturbation experiments, we added 30 meters of soil thickness to each subbasin of SRB.

The second factor of importance is soil texture (sand or clay percentages). We replaced the soil van Genuchten parameters in the SRB with those from soil classes that were randomly selected from two survey areas in the Southeast. One survey area has many map units, each of which has many soil component and horizons. We randomly selected one soil horizon from each survey area (GA603 and GA632). The soil van Genuchten parameters can be obtained by the Rosetta program. We also selected two SSURGO horizons where one had the maximum sand content (FL131) and the other one had the minimum sand content (TN081). Hence, in these experiments, the SRB basins effectively are given the same soil texture as the Coastal Plains. The characteristics of soil texture of these four SSURGO entries are shown in Table 1 (sand, silt, and clay percentages). One could note that basins on the coastal plains have much more sandy soils, and thus have high infiltration capacity.

The third factor to be analyzed was the terrain slope. We examined the difference between the slopes of the southeastern CONUS (Class #1) and SRB, which are <10% and ~30%, respectively. Thus, we implemented an experiment where the terrain slope was reduced by 80%, by changing the digital elevation data that were inputs to the data pre-processor (Shen et al., 2014) of PAWS+CLM. 80% was chosen because after this treatment, the average slopes of the SRB basins were similar to those on the coastal plains.

Besides single factor experiments, we also evaluated how multiple factors interacted to impact hydrologic fluxes. After implementing the numerical experiments, we recalculated the SSCS from each perturbed simulation. The total simulated water stored in the soil column and groundwater in the model was used as the water storage, while streamflow was extracted from the simulated daily outflow from each subbasin.

3.5 The data-centric Bayesian learning framework

The effects of the ML hypotheses can be demonstrated solely by visualizing the results from the experiments. However, as an exploratory step, here we also propose a quantitative, data-centric Bayesian framework to integrate data and the results from the modeling experiments. Essentially, the data mining provides the prior, and the numerical experiments compute a likelihood for a factor being the causal factor. The two probabilities can be integrated using the Bayes law.

Here, we define y as the observed patterns and F as the list of perturbations of the "process parameters", in other words, physical factors whose effects can be represented by perturbing our PBM. In the present example, F can take one of three values in {"soil thickness", "soil texture", “slope”$\}$. When *F* is equal to “soil thickness”, the setup of the PBM experiment is to increase soil thickness, while leaving soil texture and slope untouched. We can then identify the factors *causing* the differences in observed patterns between instances using the Bayes law:

$$P(F|y) = \frac{L(y|F)P(F)}{P(y)} \tag{1}$$

where $P(F)$ is the prior probability of the process parameters being the cause of the observed differences between instances, to be obtained from the pure data-driven analysis (more below), $L(y|F)$ is the likelihood that, after making the process perturbations in F, the differences in patterns in y in observed, $P(y) = \int L(y|F)P(F)dF$ is the marginalized probability, and $P(F|y)$ is then the probability that, given the evidence with the model experiments, F is the causal factor for the observed differences. In the Bayesian analysis here, we only consider the top three individual factors as potential values for *F,* and do not consider parameter interactions.

More specifically for this case, we start from basins that are by default of SSCS class #2 and #3 in the SRB, and ask whether a change in one of the physical factors could turn them into class #1. Therefore, $P(F)$ is the prior probability of each process perturbation, and was calculated as the frequency that *F* appears as the first level split in the RF model trained to predict the distance to the class center #1; $L(y|F)$ is the likelihood function for the perturbed model to produce class #1 basins. This likelihood was assessed using a Gaussian Mixture Model (GMM), which is a generalization from K-means clustering. Instead of predicting one class membership, the GMM generates a fuzzy membership for all classes. Our GMM used the clustering results of FS17, including the clusters’ centroids, clusters’ covariances, and the fraction of data points belonging to each class (more details of the GMM are in Appendix A). The marginalized probability, $P(y)$, was computed by integration.

The definitions of $P(F)$, which uses model visit frequency, may seem unestablished. However, in the world-shocking event where AlphaGo defeated the Go world champion, the algorithm selected the most visited move during its Monte Carlo tree search as its actual action (Silver et al., 2016). Their choice, also reliant on model visit frequency, also seemed informal, but it performed marvelously well. Our choices were based on the current best tool we have given the overall objective of this paper.

4. Results and discussions

In this section, we first show the limitations of CART and ML in general, and present multiple competing hypotheses from ML. After demonstrating the performance of the PAWS+CLM model for the Susquehanna River basin, we show results from the perturbation experiments. Finally, we put those results in the exploratory Bayesian framework and examine its usefulness.

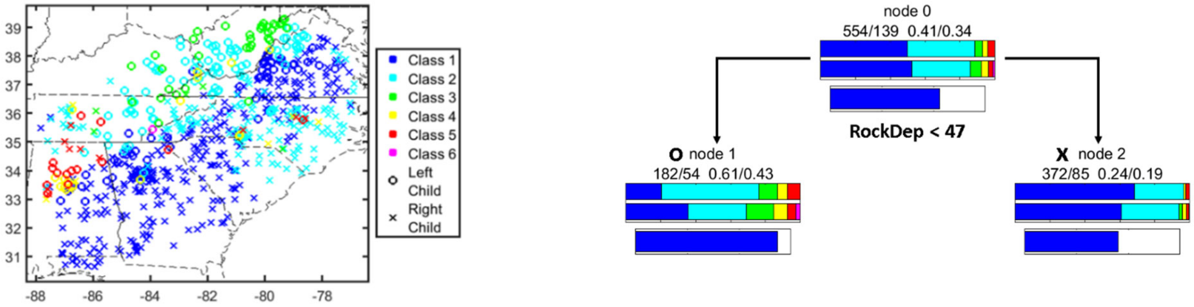
4.1 The robustness of CART and the competing hypotheses

While soil thickness was the most frequent factor that can predict the SSCS difference between class #1 and class #3 basins (Figure 4a-b), we found that soil texture (Figure 4c-d display the sand result), and terrain slope (Figure 4e-f) are competing hypotheses. The CART experiments with 20 different random seeds showed that there is a 75% chance that RockDep was selected as the top-level split, followed by Sand and then Slope. From the RF modeling, RockDep, Sand, and Slope have 21%, 17%, and 2% chances to be selected as the top-level split, respectively, with the other remaining chances mostly taken by climatic variables. The performance of these alternative models are weaker than soil thickness, but the difference, especially between soil thickness and soil texture, was not big enough to warrant confident rejection. These competing hypotheses exist because terrain slope, soil texture, and depth to bedrock covary in space. As we go from the

Appalachian mountain ranges (Appalachian Plateau, Piedmont, Valley and Ridge) to the coastal plains, simultaneously the terrain flattens, the soil texture becomes more sandy, and the soil thickness increases substantially.

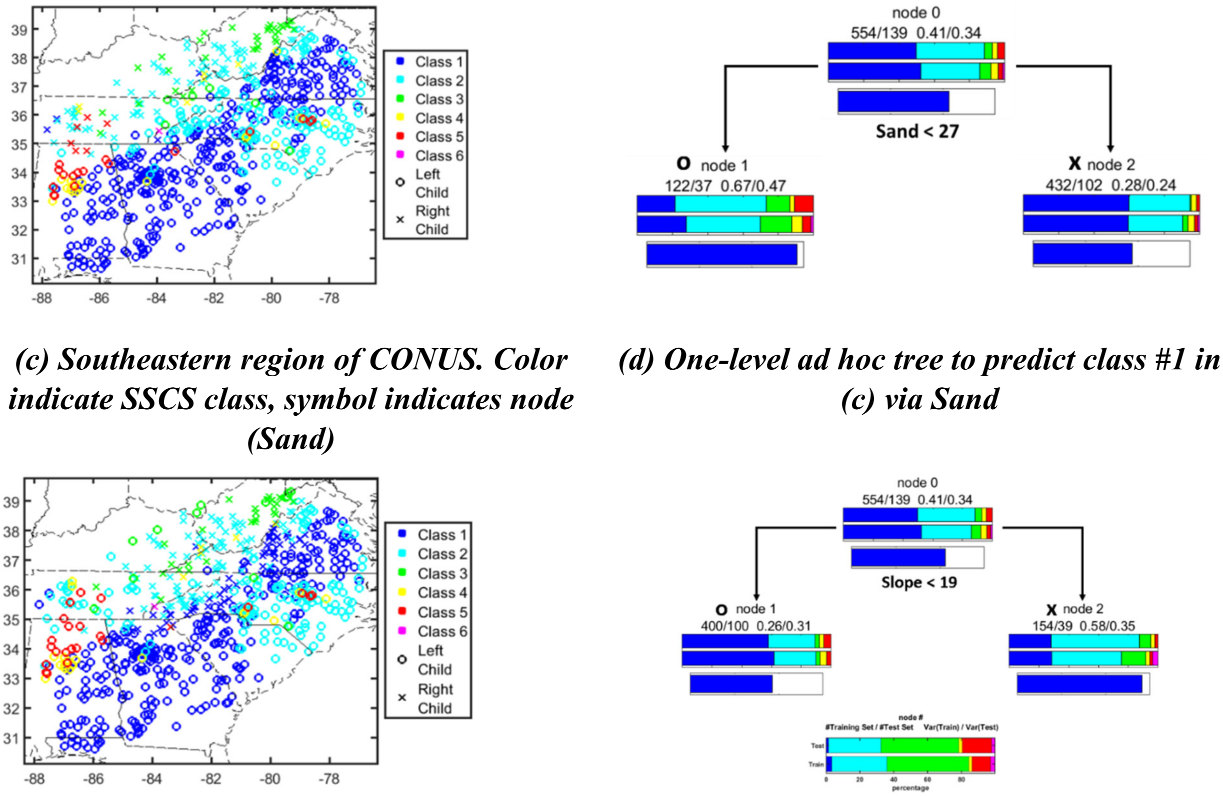
Besides random seeds, we also ran experiments with reduced training data points to examine the robustness of CART. We found that the frequency of the first-level criterion of the classification tree changed significantly when we randomly removed ~22% of the data. Moreover, in the extreme case, if we purposefully removed as few as 7 data points with the lowest sand percentages out of 693 total data points, the most important variable would change from ‘RockDep’ to ‘Sand’.

These results all suggest that the CART analysis is not robust. CART is indeed problematic; however, this is not just an issue with CART, but more generically an issue with the statistical power of the data. It can be argued that there is not enough statistical power in the data to differentiate between the causal and the coincidental factors. Geoscientists are opportunistic in the sense that we can only examine basins with the combinations of land use, geology, soil texture, and slope that naturally exist in the world and have been, or are, under study. It is not be hard to imagine missing some critical combinations which would lead to erroneous conclusions.



(a) Southeastern region of CONUS. Color indicate SSCS class, symbol indicates node (RockDep)

(b) One-level ad hoc tree to predict class #1 in (a) via RockDep



(c) Southeastern region of CONUS. Color indicate SSCS class, symbol indicates node (Sand)

(d) One-level ad hoc tree to predict class #1 in (c) via Sand

(e) Southeastern region of CONUS. Color indicate SSCS class, symbol indicates node (Slope)

(f) One-level ad hoc tree to predict class #1 in (e) via Slope

Figure 4. A one-level classification tree model picks up soil thickness (RockDep) as the main difference between two types of storage-streamflow correlation patterns (From FS17).

More importantly, from these results, we extracted three factors that are treated as competing hypotheses that explains the main difference in SSCS between the Appalachian basins and their Southeast neighbors: soil thickness (RockDep), soil texture (Sand, Silt, or Clay), and terrain. Other basin parameters such as soil bulk density and land use have very low importance and can be ignored in later analysis. We then implemented changes in these factors in the process-based model to examine their impacts on the SSCS.

4.2 Performance of the physically-based model

The daily observed USGS streamflow and simulated flow for a period of 18 years (2000-2017) were compared in Figure 5. The model had decent performance for streamflow simulation, especially within the baseflow and low flow periods (Figure 5), and captures the long-term streamflow pattern as well as some extreme high flows. The Nash-Sutcliffe model efficiency coefficient is not as high as in some of our previous applications (e.g. (Shen and Phanikumar, 2010; Shen et al., 2014; Niu et al., 2017)), due to the compromise in the 4 subbasins' parameter calibration. While the largest dam on the Susquehanna River, the Conowingo Dam, is downstream from our gage, there are other smaller dams in the basin that could have contributed to the mismatch. In addition, our experiences have indicated that NLDAS precipitation often underestimates the peak storms, leading to an under-estimation of peaks. As the main focus of the paper is not streamflow prediction, our calibration of the model is not extensive.

4.3 Testing competing hypotheses

It is easy to observe the impacts of soil thickness on the SSCS curves extracted from the default and perturbed simulations (Figure). On this figure, we colored experiments by whether they do have thicker soil implemented (adding 30 m to the soil thickness, shown in blue) or do not (shown in red). All four basins have similar patterns. The default SSCS (red x) curves are similar to SSCS classes #2 and #3 of FS17 (except the trough band of PIN, which is similar to Class #4), in that they have low correlations in peak-storage-low-flow bands, medium correlations in peak-storage-high-flow bands, and low correlations in trough-storage bands. These patterns all indicate a limited system memory; the water storage in the wet season has no impact on baseflow later in the water year. When we increased the soil thickness, the correlations in peak-storage-low-flow increased substantially, indicating that the annual-scale system memory had been enhanced. Except for the OCT subbasin, there is a clear separation between the red and blue points.

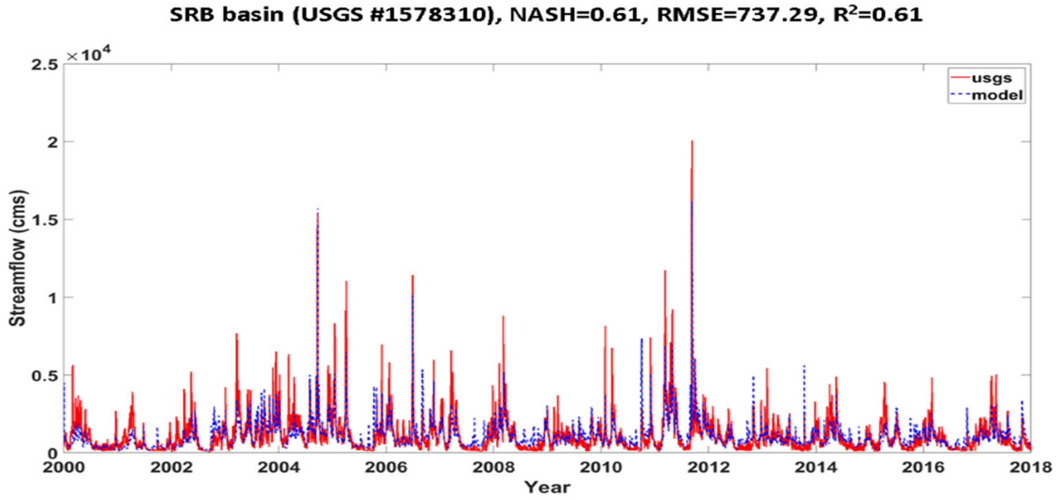
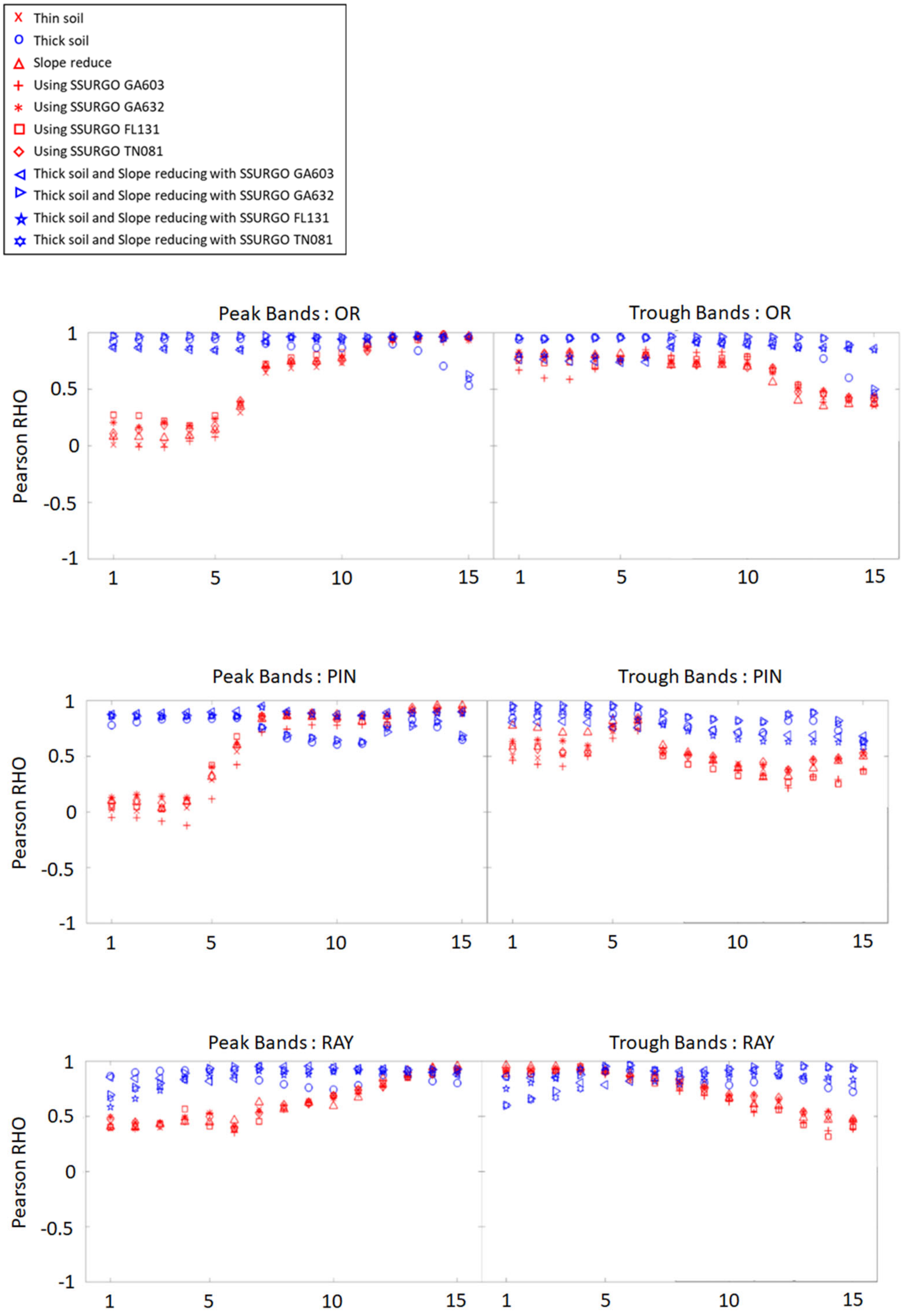

SRB basin (USGS #1578310), NASH=0.61, RMSE=737.29, R^2=0.61

Figure 5. Model streamflow simulation of whole SRB streamflow simulation. The red solid line indicates the USGS measured streamflow, and the blue dashed line indicates the model's simulated flow

On the other hand, when soil texture was modified from the default (red x) into those from the Southeast (red plus, asterisk, square, and diamond), SSCS barely fluctuated, and results based on these southeastern soil textures were clustered closely with the default simulation. We could see that soil texture has a small impact: FL131 (red square) appears to encourage higher correlations across the spectrum as compared to the others. The notable soil texture characteristics were that GA603 had a high sand percentage (most were higher than 70%); GA632 had high sand and high silt percentages (summation of both were higher than 70%); FL131 was high in sand percentage (most were higher than 80%); and TN081 was high in silt percentage (most were higher than 50%).

However, the magnitude of the impact of soil texture was not comparable to that of the soil thickness. According to the likelihood value calculated by the GMM, with all default parameters, OR belongs to Class #2 (highest probability, almost 1) and PIN belongs to Class #2 with a likelihood of 0.75 (Figure a,b). In contrast, all experiments with “thick soil” had SSCS class #1. Some parameter interaction can be observed, but its effects were minor compared to the impact of soil thickness.

From the experiments where we replaced forcing data in the SRB with those from the coastal plains, we found the impacts of climate on SSCS classes (or GMM likelihoods) to be small (data not shown here). In fact, going from Appalachia in the North to the coastal plains in the South, we saw a lower fraction of precipitation as snow, which should have reduced storage-streamflow relationships, but this effect ran counter to the observation of higher correlations between storage and streamflow in the south. Apparently, the effects of climatic variables were not as strong as the physical basin parameters, and were also coincidental factors. Hence, they were not further examined.



x Thin soil
o Thick soil
△ Slope reduce
+ Using SSURGO GA603
* Using SSURGO GA632
□ Using SSURGO FL131
◇ Using SSURGO TN081
◁ Thick soil and Slope reducing with SSURGO GA603
▷ Thick soil and Slope reducing with SSURGO GA632
★ Thick soil and Slope reducing with SSURGO FL131
✡ Thick soil and Slope reducing with SSURGO TN081
Peak Bands : OR
Trough Bands : OR
Peak Bands : PIN
Trough Bands : PIN
Peak Bands : RAY
Trough Bands : RAY
Pearson RHO
1
0.5
0
-0.5
-1
1
5
10
15

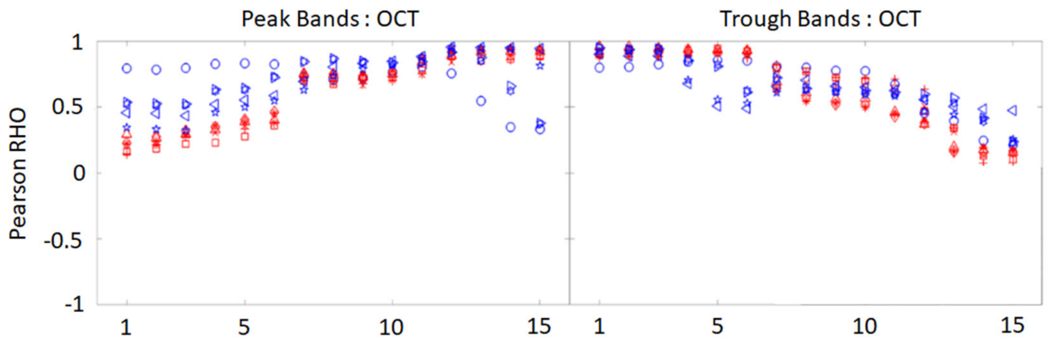
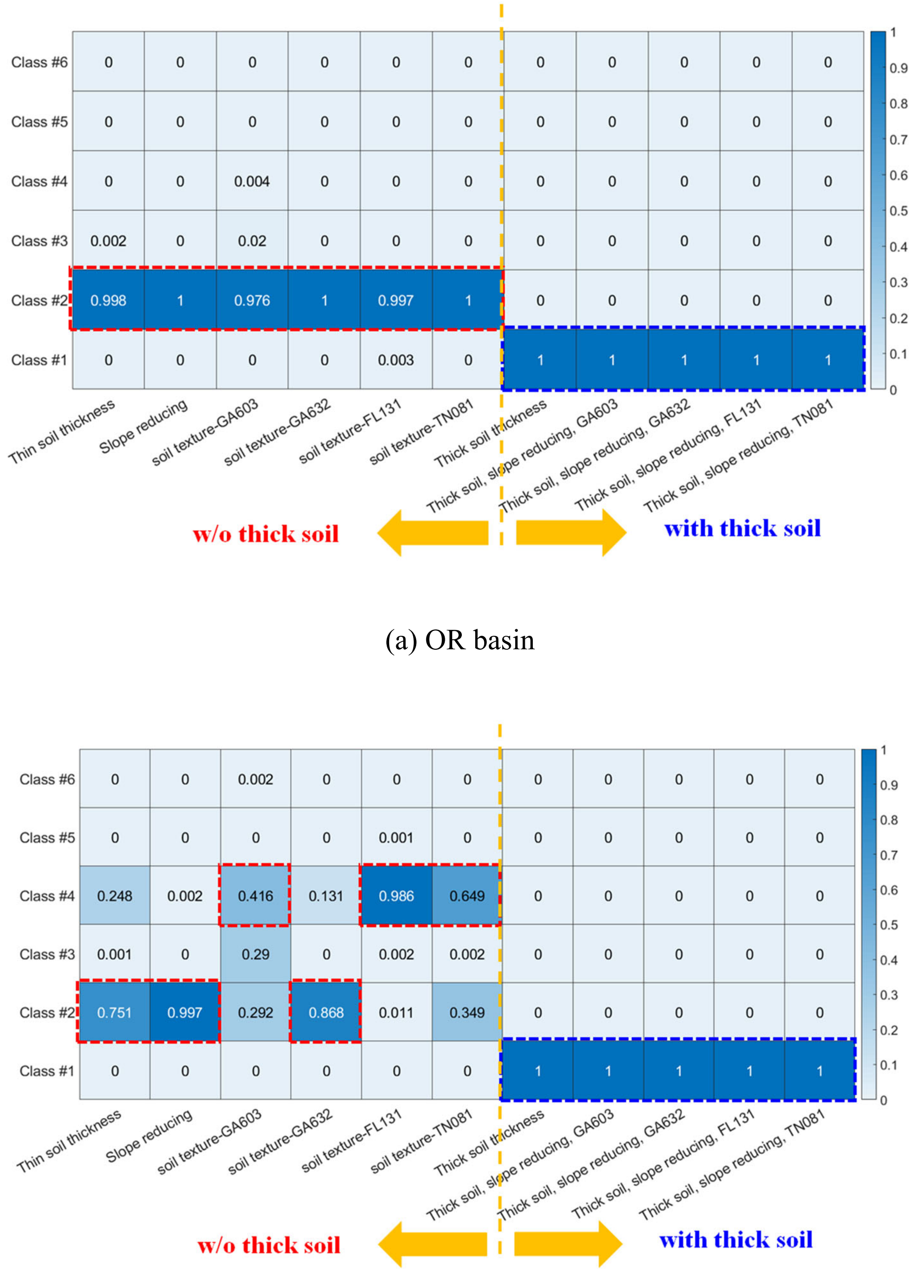


Figure 6. SSCS extracted from the numerical experiments. “Thin soil” is the default simulation with SRB-default parameters.



	Thin soil thickness	Slope reducing	soil texture-GA603	soil texture-GA632	soil texture-FL131	soil texture-TN081	Thick soil thickness	Thick soil, slope reducing, GA603	Thick soil, slope reducing, GA632	Thick soil, slope reducing, FL131	Thick soil, slope reducing, TN081
Class #6	0	0	0	0	0	0	0	0	0	0	0
Class #5	0	0	0	0	0	0	0	0	0	0	0
Class #4	0	0	0.004	0	0	0	0	0	0	0	0
Class #3	0.002	0	0.02	0	0	0	0	0	0	0	0
Class #2	0.998	1	0.976	1	0.997	1	0	0	0	0	0
Class #1	0	0	0	0	0.003	0	1	1	1	1	1

(a) OR basin

	Thin soil thickness	Slope reducing	soil texture-GA603	soil texture-GA632	soil texture-FL131	soil texture-TN081	Thick soil thickness	Thick soil, slope reducing, GA603	Thick soil, slope reducing, GA632	Thick soil, slope reducing, FL131	Thick soil, slope reducing, TN081
Class #6	0	0	0.002	0	0	0	0	0	0	0	0
Class #5	0	0	0	0	0.001	0	0	0	0	0	0
Class #4	0.248	0.002	0.416	0.131	0.986	0.649	0	0	0	0	0
Class #3	0.001	0	0.29	0	0.002	0.002	0	0	0	0	0
Class #2	0.751	0.997	0.292	0.868	0.011	0.349	0	0	0	0	0
Class #1	0	0	0	0	0	0	1	1	1	1	1

(b) PIN basin

Figure 7. The likelihood function L(y|F) as calculated by GMM in different PAWS+CLM experiments. Here, we only show the OR and PIN subbasins, but the other 2 subbasins have similar results (Appendix B).

Table 1 The characteristics of alteration of soil texture

Soil Category	Sand percentage (%)	Silt percentage (%)	Clay percentage (%)
GA603	86	4	10
GA632	43	40	17
FL131	85	10	5
TN081	21	55	25
SRB Average	32.8	51.7	15.5

4.4 The data-centric Bayesian inference results

According to the Bayesian inference framework in Equation 1, the soil thickness factor had the highest posterior probability (Table 2). Although soil texture also had a prior that was comparable to that of soil thickness, experiments that only perturbed soil had very low likelihood functions, lowering its posterior to almost zero. Terrain slope had a lower prior (although it was higher than other physical factors which were examined but not mentioned here), and its likelihood was also low, indicating that it was only a coincidental factor, not causal.

These results unequivocally support soil thickness as the causal factor of SSCS differences between Appalachian basins and those on the Southeastern coastal plains, whereas soil texture and slope were merely coincidental factors. It is notable that the PBM was needed to break the practical tie between the priors of soil texture and soil thickness. Fom these results, we can conclude that in general, systems with large soil thickness have longer memory, allowing water from the recharge season to accumulate, which thus impacts the baseflow in the hot summers. Although more sandy soil could allow for more infiltration and hence mildly boost storage-streamflow correlations, its impact was apparently not comparable to that of soil thickness. This contrast was automatically highlighted by the Bayesian framework proposed here.

Table 2. Calculations of the data-centric Bayesian inference framework for three factors. The remaining P(F) was mostly taken by climatic variables

OR basin		$P(F)$	$L(y\|F)$ ***(Class 1)***	$P(y)$	$P(F\|y)$ ***(Class 1)***
Thickness	30m addition	0.21 (P1)	0.99999 (L1)	0.21012 (P1*L1+P2*L2+P3*L3)	***1.00***
Slope	80% reduction	0.02 (P2)	0.00001 (L2)		0.00
Soil texture	Different SSURGO	0.17 (P3)	0.00070 (L3)		0.00
PIN basin		$P(F)$	$L(y\|F)$ ***(Class 1)***	$P(y)$	$P(F\|y)$ ***(Class 1)***
Thickness	30m addition	0.21	0.99997	0.21001	***1.00***
Slope	80% reduction	0.02	0.00020		0.00
Soil texture	Different SSURGO	0.17	0.00004		0.00

4.5 Further discussion

In this case study, ML allowed us to focus on only three factors prior to running any numerical experiments. Not only does this provide savings of computational power and time, but also means that we need to objectively confront our PBMs with the identified ML hypotheses. If the PBM at hand is not able to represent the effects of these factors, one needs to take note and either refine the PBM or select a different one. Because of the target, inputs, training data, and other aspects of ML still needing to be defined by humans, it is not unbiased, and fairness in artificial intelligence is a big topic (Zou and Schiebinger, 2018). However, as long as the initial ML problem is posed inclusively, ML can be relatively impartial compared to only using one PBM and starting only from expert-conceived hypotheses. The PBM was also critically important here, allowing us to study causal relationships and nuances of parameter interactions, where data may not be sufficient for complete analysis via ML.

The proposed framework is very different from that of physics-guided machine learning (PGML) (Ganguly et al., 2014; Jia et al., 2019) in that it utilizes established PBMs, which are valuable assets which the geoscience community has accumulated over the past decades, as the backbone of the analysis, whereas PGML relies on ML algorithms as the backbone. While one can easily encode simple principles such as mass and energy conservation in the loss function for PGML, it will be quite difficult to similarly express the complex physical processes and cross-domain interactions encoded in complex PBMs. Another PGML method is to pre-train network with outputs from the PBM; in the future it will certainly be interesting to compare these methods in terms of their capability and clarity of finding explanations.

The proposed data-centric Bayesian framework is raised here for the first time, and is thus only exploratory. It requires the definition of a prior (from ML), a proper PBM, a likelihood function (calculated by the GMM), and a marginalization strategy. Upon proper definition of the prior and likelihood functions, this framework can be autonomously executed. The prior is obtained from purely data analysis of GRACE and streamflow data while the posterior mostly depends on the assumed model dynamics which was built from physical laws such as Richards equation, diffusive flow equation and ecosystem equations. Each one of these choices can have alternatives, and may involve arbitrary decisions that lead to debates. We fully recognize that the choices we made could be improved in the future. However, our goal here was to highlight the value of both PBM and ML, and to inspire exploration into the diverse ways that both approaches can be coupled together for the advancement of knowledge.

The CART model was used in this study as it is most easily interpretable. Other more powerful forms of ML, e.g., time series deep learning (Fang et al., 2017, 2018), have emerged and are transforming many disciplines including hydrology (Shen, 2018). However, they are not easily interpretable and does not easily lend to hypothesis testing with the PBM.

5. Conclusions

Here we have proposed a Bayesian framework that combines machine learning and process-based modeling to overcome the limitations of both approaches. In this framework, machine learning is first used to generate competing hypotheses that are consistent with existing data. These hypotheses are subsequently implemented as perturbed process-based model simulations, which help to distinguish between causal and coincidental factors. This framework can be executed by a program and could be regarded as giving PBMs to machine learning as diagnosis tools. ML has its limitations regarding robustness, the statistical power of limited data, and causal reasoning, but it allows us to rapidly focus on several competing hypotheses and limit our subjective bias when choosing a model.

We tested the framework using the example of inferring the physical factor that controls storage-streamflow correlation behaviors across the gradients from Appalachia to the coastal plains. Although machine learning suggested that soil thickness and soil texture have similar prior probabilities of being the causal factor, the PBM experiments unequivocally supported soil thickness. This example highlights the value of the PBM in the era of big data, and promotes an alternative ML-PBM integration methodology to physics-guided machine learning, as it works with complicated, established PBMs.

Acknowledgements

This work was supported by the Office of Biological and Environmental Research of the U.S. Department of Energy under contract DE-SC0016605. Streamflow data can be downloaded from the U.S. Geological Survey Water Data for the Nation website (http://dx.doi.org/10.5066/F7P55KJN). GRACE TWSA data can be downloaded from GRACE monthly mass grids (https://grace.jpl.nasa.gov/data/get-data/). CART, RF, and GMM codes can be downloaded from Scikit-learn (https://scikit-learn.org/stable/). This article has been deposited

in self-archiving preprint server with non-exclusive rights while it is under review: https://doi.org/10.1002/essoar.10503650.1 (Tsai et al., 2020).

Appendix A. Details of Gaussian Mixture Models

Gaussian mixture models (GMM) are probabilistic models which fit a mixture –of Gaussian models to the training data. The role of a GMM is similar to that of K-mean clustering in that it predicts the class membership of an instance based on observable inputs (in this case SSCS) and the location of class centers in the input space, but the main difference is that a GMM produces a probability for each class. In probability-based clustering, it is assumed that the data are generated by a mixture of underlying probability distributions in which each component represents a different group or cluster (Fraley & Raftery, 1998). Given observations $x = (x_1, x_2, \dots, x_n)$, let $f_k(x_i|\theta_k)$ be the density of an observation x_i from the *k*th class, where θ_k are the corresponding parameters, and let *G* be the number of classes in the mixture. When $f_k(x_i|\theta_k)$ is multivariate normal (Gaussian), we call it a Gaussian mixture model. The model for the composite of the clusters is usually formulated in one of two ways. The classification likelihood approach maximizes

$$L_C(\theta_1, \dots, \theta_G; \gamma_1, \dots, \gamma_n|x) = \prod_{i=1}^{n} f_{\gamma_i}(x_i|\theta_{\gamma_i}) \qquad (2)$$

where γ_i are discrete values labelling the classification: $\gamma_i = k$ if x_i belongs to the *k*th component. The mixture likelihood approach maximizes

$$L_M(\theta_1, \dots, \theta_G; \tau_1, \dots, \tau_G|x) = \prod_{i=1}^{n} \sum_{k=1}^{G} \tau_k f_k(x_i|\theta_k) \qquad (3)$$

where τ_k is the probability that an observation belongs to the *k*th component: $(\tau_k \geq 0; \sum_{k=1}^{G} \tau_k = 1)$.

A more detailed description of GMM can be found in Fraley & Raftery (1998) and Huang et al. (2005). One of the important issues of GMM is model order/number. Determining model order is a critical, but difficult factor in training a GMM. However, according to FS17, the suitable classes for SSCS of CONUS (6 classes) have been identified. In this study, each class central in FS17 was selected as each Gaussian mixture model.

Appendix B. The likelihood function by GMM in different PAWS+CLM experiments.

Basin: Ray	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
Thin soil thickness	0.276	**0.724**	0.000	0.000	0.000	0.000
Slope reducing	0.258	**0.742**	0.000	0.000	0.000	0.000
Soil texture - GA603	0.006	**0.994**	0.000	0.000	0.000	0.000
Soil texture - GA632	0.083	**0.917**	0.000	0.000	0.000	0.000
Soil texture - FL131	0.023	**0.977**	0.000	0.000	0.000	0.000
Soil texture - TN081	0.357	**0.643**	0.000	0.000	0.000	0.000
Thick soil thickness	**1.000**	0.000	0.000	0.000	0.000	0.000
Thick soil, slope reducing, GA603	**1.000**	0.000	0.000	0.000	0.000	0.000
Thick soil, slope reducing, GA632	**1.000**	0.000	0.000	0.000	0.000	0.000
Thick soil, slope reducing, FL131	**1.000**	0.000	0.000	0.000	0.000	0.000
Thick soil, slope reducing, TN081	**1.000**	0.000	0.000	0.000	0.000	0.000

Basin: OCT	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
Thin soil thickness	0.000	**0.999**	0.000	0.000	0.000	0.000
Slope reducing	0.009	**0.991**	0.000	0.000	0.000	0.000
Soil texture - GA603	0.006	**1.000**	0.000	0.000	0.000	0.000
Soil texture - GA632	0.112	**0.886**	0.000	0.000	0.000	0.000
Soil texture - FL131	0.000	**1.000**	0.000	0.000	0.000	0.000
Soil texture - TN081	0.036	**0.959**	0.000	0.000	0.000	0.000
Thick soil thickness	**1.000**	0.000	0.000	0.000	0.000	0.000
Thick soil, slope reducing, GA603	**1.000**	0.000	0.000	0.000	0.000	0.000
Thick soil, slope reducing, GA632	**1.000**	0.000	0.000	0.000	0.000	0.000
Thick soil, slope reducing, FL131	**1.000**	0.000	0.000	0.000	0.000	0.000
Thick soil, slope reducing, TN081	**1.000**	0.000	0.000	0.000	0.000	0.000

References

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Boca Raton, FL, USA: CRC Press.

Collins, W. D., Bitz, C. M., Blackmon, M. L., Bonan, G. B., Bretherton, C. S., Carton, J. A., et al. (2006). The Community Climate System Model Version 3 (CCSM3). *J. Clim.* 19, 2122–2143. doi:10.1175/JCLI3761.1.

Dickinson, R. E., Oleson, K. W., Bonan, G., Hoffman, F., Thornton, P., Vertenstein, M., et al. (2006). The Community Land Model and Its Climate Statistics as a Component of the Community Climate System Model. *J. Clim.* 19, 2302–2324. doi:10.1175/JCLI3742.1.

Dingman, S. L. (2015). *Physical hydrology*. Third. Long Grove, IL: Waveland Press.

Fang, K., Ji, X., Shen, C., Ludwig, N., Godfrey, P., Mahjabin, T., et al. (2019). Combining a land surface model with groundwater model calibration to assess the impacts of groundwater pumping in a mountainous desert basin. *Adv. Water Resour.* 130, 12–28. doi:10.1016/J.ADVWATRES.2019.05.008.

Fang, K., Pan, M., and Shen, C. (2018). The Value of SMAP for Long-Term Soil Moisture Estimation With the Help of Deep Learning. *IEEE Trans. Geosci. Remote Sens.*, 1–13. doi:10.1109/TGRS.2018.2872131.

Fang, K., and Shen, C. (2017). Full-flow-regime storage-streamflow correlation patterns provide insights into hydrologic functioning over the continental US. *Water Resour. Res.* 53, 8064–8083. doi:10.1002/2016WR020283.

Fang, K., Shen, C., Kifer, D., and Yang, X. (2017). Prolongation of SMAP to Spatio-temporally Seamless Coverage of Continental US Using a Deep Learning Neural Network. *Geophys. Res. Lett.* 44, 11030–11039. doi:10.1002/2017GL075619.

Ganguly, A. R., Kodra, E. A., Agrawal, A., Banerjee, A., Boriah, S., Chatterjee, S., et al. (2014). Toward enhanced understanding and projections of climate extremes using physics-guided data mining techniques. *Nonlinear Process. Geophys.* 21, 777–795. doi:10.5194/npg-21-777-2014.

Guber, A. K., Pachepsky, Y. A., van Genuchten, M. T., Simunek, J., Jacques, D., Nemes, A., et al. (2009). Multimodel simulation of water flow in a field soil using pedotransfer functions. *Vadose Zo. J.* 8, 1. doi:10.2136/vzj2007.0144.

Ho, T. K. (1995). Random decision forests. in *Proceeding ICDAR '95 Proceedings of the Third International Conference on Document Analysis and Recognition*.

Ji, X., Lesack, L., Melack, J. M., Wang, S., Riley, W. J., and Shen, C. (2019). Seasonal and inter-annual patterns and controls of hydrological fluxes in an Amazon floodplain lake with a surface-subsurface processes model. *Water Resour. Res.* 55, 3056–3075. doi:10.1029/2018WR023897.

Ji, X., and Shen, C. (2018). The introspective may achieve more: enhancing existing Geoscientific models with native-language structural reflection. *Comput. Geosci.* 110. doi:10.1016/j.cageo.2017.09.014.

Ji, X., Shen, C., and Riley, W. J. (2015). Temporal evolution of soil moisture statistical fractal and controls by soil texture and regional groundwater flow. *Adv. Water Resour.* 86, 155–169. doi:10.1016/j.advwatres.2015.09.027.

Jia, X., Karpatne, A., Willard, J., Steinbach, M., Read, J., Hanson, P. C., et al. (2018). Physics Guided Recurrent Neural Networks For Modeling Dynamical Systems: Application to Monitoring Water Temperature And Quality In Lakes. in *8th International Workshop on Climate Informatics* Available at: http://arxiv.org/abs/1810.02880 [Accessed August 3, 2019].

Jia, X., Willard, J., Karpatne, A., Read, J., Zwart, J., Steinbach, M., et al. (2019). Physics Guided RNNs for Modeling Dynamical Systems: A Case Study in Simulating Lake Temperature Profiles. in *Proceedings of the 2019 SIAM International Conference on Data Mining* (Philadelphia, PA: Society for Industrial and Applied Mathematics), 558–566. doi:10.1137/1.9781611975673.63.

Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., et al. (2017). Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data. *IEEE Trans.*

Knowl. Data Eng. 29, 2318–2331. doi:10.1109/TKDE.2017.2720168.

Lawrence, D. M., Oleson, K. W., Flanner, M. G., Thornton, P. E., Swenson, S. C., Lawrence, P. J., et al. (2011). Parameterization improvements and functional and structural advances in Version 4 of the Community Land Model. *J. Adv. Model. Earth Syst.* 3. doi:10.1029/2011MS000045.

Maxwell, R. M., and Condon, L. E. (2016). Connections between groundwater flow and transpiration partitioning. *Science (80-.).* 353.

May, J. (2011). On ancient Susquehanna River, flooding's a frequent fact. *Assoc. Press*. Available at: http://cumberlink.com/news/local/on-ancient-susquehanna-river-flooding-s-a-frequent-fact/article_ee769266-db2f-11e0-945d-001cc4c002e0.html [Accessed March 31, 2016].

Mitchell, K. E. (2004). The multi-institution North American Land Data Assimilation System (NLDAS): Utilizing multiple GCIP products and partners in a continental distributed hydrological modeling system. *J. Geophys. Res.* 109, D07S90. doi:10.1029/2003JD003823.

Mitchell, T. (1997). *Machine Learning*. McGraw-Hill Science/Engineering/Math.

Nash, J. E., and Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I — A discussion of principles. *J. Hydrol.* 10, 282–290. doi:10.1016/0022-1694(70)90255-6.

Nearing, G. S., Mocko, D. M., Peters-Lidard, C. D., Kumar, S. V., and Xia, Y. (2016). Benchmarking NLDAS-2 Soil Moisture and Evapotranspiration to Separate Uncertainty Contributions. *J. Hydrometeorol.* 17, 745–759. doi:10.1175/JHM-D-15-0063.1.

Niu, J., Shen, C., Chambers, J., Melack, J. M., and Riley, W. J. (2017). Interannual variation in hydrologic budgets in an Amazonian watershed with a coupled subsurface - land surface process model. *J. Hydrometeorol.* doi:10.1175/JHM-D-17-0108.1.

NRCS (2010). SSURGO Soil Survey Geographic Database. *Nat. Resour. Conserv. Serv.* Available at: http://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/?cid=nrcs142p2_053627.

Oleson, K., Lawrence, D. M., Bonan, G. B., Flanner, M., Kluzek, E., Lawrence, P., et al. (2010). Technical description of version 4.0 of the Community Land Model (CLM). Boulder, Colorado: NCAR Technical Note, NCAR/TN-478+STR.

Pelletier, J. D., Broxton, P. D., Hazenberg, P., Zeng, X., Troch, P. A., Niu, G.-Y., et al. (2016). A gridded global data set of soil, intact regolith, and sedimentary deposit thicknesses for regional and global land surface modeling. *J. Adv. Model. Earth Syst.* doi:10.1002/2015MS000526.

Poff, N. L., and Allan, J. D. (1995). Functional Organization of Stream Fish Assemblages in Relation to Hydrological Variability. *Ecology* 76, 606. doi:10.2307/1941217.

Reager, J. T., Thomas, A., Sproles, E., Rodell, M., Beaudoing, H., Li, B., et al. (2015). Assimilation of GRACE Terrestrial Water Storage Observations into a Land Surface Model for the Assessment of Regional Flood Potential. *Remote Sens.* 7, 14663–14679. doi:10.3390/rs71114663.

Reager, J. T., Thomas, B. F., and Famiglietti, J. S. (2014). River basin flood potential inferred using GRACE gravity observations at several months lead time. *Nat. Geosci.* 7, 588–592. doi:10.1038/ngeo2203.

Russell, S., and Norvig, P. (2009). *Artificial Intelligence: A Modern Approach*. 3rd ed. Upper Saddle River, NJ: Prentice Hall Available at: https://www.amazon.com/Artificial-Intelligence-Approach-ARTIFICIAL-INTELLIGENCE/dp/B008NYIYZS/ref=as_li_ss_tl?ie=UTF8&qid=1543442234&sr=8-7&keywords=Artificial+Intelligence:+A+Modern+Approach&linkCode=sl1&tag=inspiredalgor-20&linkId=d6316766f72fb3d953858b [Accessed July 17, 2019].

Schaap, M. G., Leij, F. J., and van Genuchten, M. T. (2001). Rosetta: a Computer Program for Estimating Soil Hydraulic Parameters With Hierarchical Pedotransfer Functions. *J. Hydrol.* 251, 163–176. doi:10.1016/S0022-1694(01)00466-8.

Shen, C. (2018). A trans-disciplinary review of deep learning research and its relevance for water resources scientists. *Water Resour. Res.* 54, 8558–8593. doi:10.1029/2018WR022643.

Shen, C., Laloy, E., Elshorbagy, A., Albert, A., Bales, J., Chang, F.-J., et al. (2018). HESS Opinions: Incubating deep-learning-powered hydrologic science advances as a community. *Hydrol. Earth Syst. Sci.* 22, 5639–5656. doi:10.5194/hess-22-5639-2018.

Shen, C., Niu, J., and Fang, K. (2014). Quantifying the effects of data integration algorithms on the outcomes of a subsurface–land surface processes model. *Environ. Model. Softw.* 59, 146–161. doi:10.1016/j.envsoft.2014.05.006.

Shen, C., Niu, J., and Phanikumar, M. S. (2013). Evaluating controls on coupled hydrologic and vegetation dynamics in a humid continental climate watershed using a subsurface - land surface processes model. *Water Resour. Res.* 49, 2552–2572. doi:10.1002/wrcr.20189.

Shen, C., and Phanikumar, M. S. (2010). A process-based, distributed hydrologic model based on a large-scale method for surface–subsurface coupling. *Adv. Water Resour.* 33, 1524–1541. doi:10.1016/j.advwatres.2010.09.002.

Shen, C., Riley, W. J., Smithgall, K. M., Melack, J. M., and Fang, K. (2016). The fan of influence of streams and channel feedbacks to simulated land surface water and carbon dynamics. *Water Resour. Res.* 52, 880–902. doi:10.1002/2015WR018086.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 484–489. doi:10.1038/nature16961.

Thomas, B. F., Vogel, R. M., Kroll, C. N., and Famiglietti, J. S. (2013). Estimation of the base flow recession constant under human interference. *Water Resour. Res.* 49, 7366–7379. doi:10.1002/wrcr.20532.

Tsai, W.-P., Fang, K., Ji, X., and Shen, C. (2020). Revealing causal controls of storage-streamflow relationships with a data-centric Bayesian framework combining machine learning and process-based modeling. *Earth Sp. Sci. Open Arch.* doi:10.1002/essoar.10503650.1.

Verhougstraete, M. P., Martin, S. L., Kendall, A. D., Hyndman, D. W., and Rose, J. B. (2015). Linking fecal bacteria in rivers to landscape, geochemical, and hydrologic factors and sources at the basin scale. *Proc. Natl. Acad. Sci. U. S. A.* 112, 10419–24. doi:10.1073/pnas.1415836112.

Vrugt, J. A., Gupta, H. V., Bastidas, L. A., Bouten, W., and Sorooshian, S. (2003). Effective and efficient algorithm for multiobjective optimization of hydrologic models. *Water Resour. Res.* 39. doi:10.1029/2002WR001746.

Vrugt, J. A., ter Braak, C. J. F., Diks, C. G. H., Robinson, B. A., Hyman, J. M., and Higdon, D. (2009). Accelerating Markov Chain Monte Carlo Simulation by Differential Evolution with Self-Adaptive Randomized Subspace Sampling. *Int. J. Nonlinear Sci. Numer. Simul.* 10, 273–290. Available at: http://apps.webofknowledge.com/full_record.do?product=UA&search_mode=GeneralSearch&qid=1&SID=3BCHPbdo77P8abAAe2e&page=1&doc=1 [Accessed April 13, 2013].

Yang, X., Barajas-Solano, D., Tartakovsky, G., and Tartakovsky, A. M. (2019). Physics-informed CoKriging: A Gaussian-process-regression-based multifidelity method for data-model convergence. *J. Comput. Phys.* 395, 410–431. doi:10.1016/J.JCP.2019.06.041.

Yarnal, B., Johnson, D. L., Frakes, B. J., Bowles, G. I., and Pascale, P. (1997). The flood of ’96 and its socioeconomic impacts in the Susquehanna River basin. *J. Am. Water Resour. Assoc.* 33, 1299–1312. doi:10.1111/j.1752-1688.1997.tb03554.x.

Zou, J., and Schiebinger, L. (2018). AI can be sexist and racist — it’s time to make it fair. *Nature* 559, 324–326. doi:10.1038/d41586-018-05707-8.