Can artificial intelligence and data-driven machine learning models match or even replace process-driven hydrologic models for streamflow simulation?: A case study of four watersheds with different hydro-climatic regions across the CONUS

Taereem Kim[a], Tiantian Yang[a, *], Shang Gao[a], Lujun Zhang[a], Ziyu Ding[b], Xin Wen[b], Jonathan J. Gourley[d], and Yang Hong[a,c]

[a] School of Civil Engineering and Environmental Science, University of Oklahoma, Norman 73072, United States.

[b] College of Water Conservancy and Hydropower Engineering, Hohai University, Nanjing 210098, China.

[c] School of Meteorology, University of Oklahoma, Norman, OK 73072, United States.

[d] NOAA/National Severe Storms Laboratory, Norman, OK 73072, United States.

[*] Corresponding author: Tiantian Yang (tiantian.yang@ou.edu)

1 Introduction

Hydrologic models are powerful tools for forecasting potential floods or droughts and managing surface and subsurface water resources. From the beginning of the 1850s to the present, hydrologic models have rapicly developed with the advanced numerical mathematics and computer revolutions (Chow, 1964; Singh, 2018; Singh and Woolhiser, 2002). The current development of the hydrologic model is towards integrating the Earth system of climate and weather, global atmospheric circulation, and geospatial characteristics (Brown et al., 2014; Senatore et al., 2020; Sorooshian et al., 2008; Tian et al., 2020). Our understanding of the water cycle's physical processes has greatly improved, and many new types of data are also become available to be used in hydrologic models. Along with these changes, hydrologic models have evolved from simple and conceptual models to various process-based and more complicated models. These models could be grouped into Process-based Hydrologic Models (PHMs) and and Data-driven Machine Learning Models (DMLs). The former models were originated from the classical bucket model, and they are also called conceptual, mechanistic, or process-driven models (Islam, 2011). The later models were derived from the traditional statistical analysis relying on mathematical regression, and their latest developments are towards using advanced computational intelligence (Clark et al., 2015; Grayson et al., 1992; Koutsoyiannis, 2003; Machiwal and Jha, 2012; Salas, 1980; Solomatine et al., 2008). The DMLs have begun to permeate the hydrologic community with rapid growth, leveraged by the blossom of different machine learning algorithms for classification and regression problems (Aytek et al., 2008; Liu et al., 2017; Mosavi et al., 2018; Rasouli et al., 2012; Wang et al., 2009). However, it is still unknown whether the DMLs will gradually prevail over the PHMs or vice versa, and how the traditional hydrologic models will further develop, given the distinct philosophy used in those two model groups. It is important to carry out a detailed evaluation of these two types of models and to identify which models are more reliable and powerful than others under what conditions, rather than simply developing new models or applying them separately in different study domains. Therefore, we set up a few benchmark hydrological simulation cases and comprehensively compared the pros and cons of a few popular PHMs and DMLs. Our research goals

are to identify in what conditions the PHMs or the DMLs could generate the most accurate streamflow simulation compared to observation and to investigate whether the newer DMLs have any potential in further improving the simulation accuracy of classical PHMs, and vice versa. The following introduction section reviews the employed PHMs and DMLs in this study.

The PHMs can be categorized into three types, which are the lumped models, semi-distributed models, and fully distributed models. In the present study, we used two lumped models (Sacramento Soil Moisture Accounting; SAC-SMA and Xinanjiang; XAJ) and one distributed model (Coupled Routing and Excess Storage; CREST). We include both lumped and distributed models because in prior studies, i.e., the two phases of the Distributed Model Intercomparison Project (DMIP 1&2) conducted by the National Weather Service, it was found that they sometimes outperform one another in simulating the streamflow and there were no major differences in their performances throughout many study cases (Grayson et al., 1992; Khakbaz et al., 2012; Koren et al., 2004). The SAC-SMA model has been well-recognized and widely used in both operational agencies and research communities as a key component of the National Weather Service River Forecast System (NWSRFS) for rainfall-runoff modeling (Behrangi et al., 2011; Boyle et al., 2001; Chu et al., 2010; Sorooshian et al., 1993). Numerous studies have applied the SAC-SMA model for flood forecasting, soil properties studies, and baseflow generation around the world (Abdulla et al., 1999; Ajami et al., 2004; Buchtele et al., 1996; Hogue et al., 2006; Hogue et al., 2000; Moreda et al., 2006). Another lumped model in this study is the XAJ model. The XAJ model has been widely used for rainfall-runoff simulation, flood forecasting, and water resources planning and management in large-scale humid and semi-humid regions, because it requires fewer forcing data to execute as compared to other pumped hydrologic models (Lu and Li, 2014; Ren-Jun, 1992; Xu et al., 2013; Zeng et al., 2018; Zhijia et al., 2013). Several studies have used these two lumped models simultaneously to forecast streamflow and achieved satisfactory simulation results under different climate conditions (Hao et al., 2018; Huang et al., 2016; Huo et al., 2019). While, other studies argued that these lumped models were limited in simulating the nonlinear process, and they only performed well in semi-humid and humid regions without low transferability to other

climate regions (Hu et al., 2005; Huang et al., 2016; Yao et al., 2009). However, the lumped hydrologic models remain prevailing in the water community because of their proven effectiveness and robustness through the years. The CREST model is a distributed hydrologic model. The key concept of the distributed hydrologic models are to divide the watershed into smaller hydrologic response units and to build individual and smaller lumped models for each hydrologic unit. All hydrologic response units are connected by mass-balance equations, and the water is routed through all meshed units and then become the total watershed discharge. In the CREST model, four excess storage reservoirs represent the interception by the vegetation canopy and subsurface water storage in one underlying soil layer (Wang et al., 2011). We chose the CREST model because it was adopted as an operational tool across the US NWS for flash flood forecasting by local NWS Forecast Offices in the Flooded Locations and Simulated Hydrographs Project (Gourley et al., 2017). With the convenience of coupling remotly sensed data with distributed hydrologic models, the number of applications of the CREST model is increasing in recent years (Kan et al., 2017; Li et al., 2018; Shen et al., 2017).

Similar to the PHMs, the DMLs have also been widely used to solve various classification and regression problems in hydrologic sciences. The DMLs can identify the statistical relationship between input and output data without the explicit requirement of users to know of the physical processes (Reichstein et al., 2019; Solomatine and Ostfeld, 2008). In some recent studies, researchers have used the DMLs to simulate complicated hydrologic processes, and some studies have achieved better performances than traditional process-based hydrologic models (Chaney et al., 2018; Chaney et al., 2016; Ham et al., 2019; Kim et al., 2019; Shen, 2018; Zhao et al., 2019). In the present study, two popular machine learning algorithms, namely the Artificial Neural Networks (ANN) and Long Short Term Memory (LSTM), are implemented to simulate the watershed discharge and compared with the PHMs (i.e., the SAC-SMA, XAJ, and the CREST models). The ANN model showed superior performance in hydrologic simulation under complex geophysical processes with the growing popularity from the literature (Kim et al., 2020; Yang et al., 2017b). It can link the climate information without requiring explicit interpretation, which the PHMs

could not achieve using the micro-scale mass-balance governing equations (Abbot and Marohasy, 2012; Aksoy and Dahamsheh, 2009; Azadi and Sepaskhah, 2012; Dahamsheh and Aksoy, 2009; French et al., 1992; Hung et al., 2009; Kim et al., 2019). Beyond the ANN model, the Recurrent Neural Networks (RNN) is another ANN class, where connections between nodes are designed to exhibit temporal dynamic behavior by processing the inputs in its sequential order (Kratzert et al., 2018; Rumelhart et al., 1986; Rumelhart et al., 1994). Unlike the traditional ANN model, the RNN model has a memory that remembers some information about a sequence, which can greatly improve the predictive performance if the inputs and outputs are correlated. Most recently, the LSTM model, a class of the RNN model, has gained lots of popularity in hydrological sciences, and has been successfully applied to solve hydrological and environmental forecasting problems (Akbari Asanjan et al., 2018; Kumar et al., 2019; Srivastava and Lessmann, 2018).

The present study also builds on many existing studies related to streamflow simulation by comparing hydrologic model performance. Previous studies have preliminarily compared the model simulation performance between PHMs and DMLs (Daliakopoulos and Tsanis, 2016; Hsu et al., 1995; Ju et al., 2009; Rauf and Ghumman, 2018; Rezaeianzadeh et al., 2013; Roodsari et al., 2019; Srivastava et al., 2006; Tokar and Markus, 2000; Wang et al., 2017). One common agreement in these studies is that each model, PHM or DMLs, may differ model performance depending on the hydrological and climate condition (such as climate, weather, terrain, and soil), data availability, and simulation objectives in the study regions. Besides, there are many discussions among hydrologists, in which concerns are raised about the lack of physical constraints and formulation of the water routing dynamics in all DMLs, though there are cases that DMLs outperform the PHMs using a flexible set of information as inputs (Kratzert et al., 2019; Sellars, 2018). Following this discussion, the designed experiments in this study will answer the questions that (1) whether the DMLs and PHMs can supplement each other at a particular simulation condition such as season or flow regimes in our study cases, and (2) how water managers could acquire a deeper understanding of the pros and cons of both PHMs and DMLs and develop more accurate streamflow simulation by taking

advantage of both types of models. To better understand the performances of the PHMs to the DMLs in simulating streamflow, this study compares a few popular PHMs and DMLs over four watersheds across the Continental US (CONUS) that are associated with different climate and regional conditions. Based on the simulation results, we further examined how the model performance will vary based on different seasonalities and magnitude of flow regimes and explore whether the PHMs and DMLs could be interchangeable under certain conditions.

The rest of this paper is organized as follows. Section 2 present the methodologies applied in this study, including the three PHMs and two DMLs. Section 3 provides detailed information on study basins and datasets. Section 4 summarizes daily streamflow simulation results at four watersheds in the CONUS. Section 5 provides the result analysis and discussion, and section 6 summarizes our conclusions and recommendations for the development of PHMs and DMLs.

2. Methodology

Figure 1 shows the conceptual comparison between PHMs (left side) and DMLs (right side). Both PHMs and DMLs follow a general hydrologic modeling framework, while there are some similarities and differences at each stage during the process. First, both PHMs and DMLs take the same forcing data as inputs and historical observation to improve the model performance. This process in the PHMs is called calibration, while in the DMLs is termed model training. The PHMs is based on mass balance equations and the governing equations, which are the key mathematical abstractions for representing the water flow in either the entire watershed or a hydrologic response unit. Various PHMs use different mathematical equations to describe the hydrologic processes, and each mathematical equation defines corresponding parameters to mimic the physical dynamics of water flows. The model parameters used in the PHMs have an important role in making sure the model outputs match the observation, and they are updated manually or automatically over the calibration period.

In contrast to PHMs, the DMLs purely relies on the intrinsic relationship between model input and output data. Statistical fitting is the primary way to establish the data relationship instead of defining a physically formulated mass balance equation of water flows. The DMLs also require initial weights to structure neural networks (i.e., layers, number of nodes, and activation functions), which are the parameters to regulate the statistical fitting function. Since DMLs try to establish the statistical relationship between input and output data, they need a large number of data records to train a model. The DMLs training is a trial-and-error process to find the minimal error terms between model simulation and observation. This process is similar to the manual or automatic parameter calibration process in the PHMs. The DMLs training always leverages optimization algorithms to tune the hyperparameters or connection weights, similar to that used in PHMs.

(Figure 1 around here)

2.1 Process-based Hydrologic Models (PHMs)

2.1.1 SAC-SMA model

The SAC-SMA model is a classical, lumped, conceptual hydrologic model originally introduced by Burnash et al. (1973). The SAC-SMA model represents the soil column with upper and lower zones of multiple storages (Burnash, 1995). The conceptual diagram of the SAC-SMA model is shown in Figure 2 (Gan et al., 2014). The SAC-SMA model takes two inputs, i.e., the mean areal precipitation and potential evapotranspiration (PET), and incorporates a two-story structure of soil moistures: upper and lower zones. Each layer of the soil water storage is divided into tension and free water storages that interact to generate soil moisture states and five runoff components at each simulation time step. The lower zone's free water storage is also divided into supplemental and primary water storage to consider the fast and slow groundwater flows. Partitioning of rainfall into surface runoff and infiltration into the lower zone storages depends on the available storage of tension and free water in the upper zone. The amount of rainfall that

161 exceeds the upper zone tension water capacity (UZTWM) becomes the excess rainfall. The amount of
162 rainfall that exceeds the upper zone free water capacity (UZFWM) becomes the surface runoff. Percolations
163 between the upper and lower zones are described by two parameters: the maximum rate of percolation
164 (ZPERC) and an exponent value (REXP). A fractional value (PFREE) is further used to split the percolated
165 water into lower zone tension and free water storage. All three free water buckets from upper and lower
166 zones have corresponding depletion coefficients (UZK, LZSK, LZPK), which describe the linear
167 relationship between the runoff and their water storage.

168 The SAC-SMA model has 16 parameters in total. Among these, 3 out of 16 parameters are not
169 tunable, according to Brazil (1989) and Peck (1976). These three parameters are (1) the percentage of the
170 lower zone free water that cannot be transferred to lower zone tension water (RSERV); (2) the riverside
171 vegetation area (RIVA); and (3) the ratio of non-channel baseflow to channel base flow (SIDE). The
172 remaining 13 parameters are tunable. Detailed information about the SAC-SMA model parameters is listed
173 in Table 1.

174 (Figure 2 around here)

175 (Table 1 around here)

176 ### 2.1.2 XinAnJiang (XAJ) model

177 The XAJ model is a conceptual lumped hydrologic model developed by Zhao (1980). The XAJ
178 model has been shown effective in simulating the hydrological processes in humid and semi-humid regions.
179 The conceptual diagram of the XAJ model is shown in Figure 3. The XAJ model has 15 tunable parameters.
180 These 15 parameters can be grouped as shown in Table 2. Similar to the SAC-SMA, the model takes two
181 inputs, i.e., the mean precipitation and evapotranspiration, and includes four different strategies to process
182 the hydrological variables, namely, (1) Evapotranspiration: using a three-layer soil moisture model with
183 upper, lower layers and deep layers; (2) Runoff production: using watershed stored-full runoff theory,

184 which means that runoff production occurs on repletion of storage to assumed capacity values of the basin;
185 (3) Separation of runoff components: dividing the total runoff into the surface runoff, interflow, and
186 groundwater; and (4) Flow concentration: using linear reservoir model. For specific details, please see the
187 reference (Zhao, 1992).

188 (Figure 3 around here)

189 (Table 2 around here)

190 ### 2.1.3 Coupled Routing and Excess Storage (CREST) model

191 The CREST model is a fully distributed hydrological model. The CREST model features a grid-
192 based water balance component coupled with the kinematic wave water routing model. It applies from
193 small- to medium-sized basins at very high-resolutions with cost-effective computational efficiency (Shen
194 et al., 2017; Xue et al., 2013). The CREST model simulates the spatial and temporal variation of the land
195 surface and subsurface water fluxes and storages on a regular grid with a user-defined grid cell resolution
196 (Wang et al., 2011). Figure 4 shows a conceptual diagram of the CREST model, showing how CREST
197 operates in each cell. The primary two inputs for a cell are precipitation and evapotranspiration. The rainfall-
198 runoff generation of CREST model is simulated by the interception in the canopy layer. Specifically, the
199 variable infiltration curve separates the precipitation reaching the soil surface into excess rain (R) and
200 infiltration water, runoff generation partitions the sub-grid routing, routing of overland, channel, and
201 subsurface components downstream, and evapotranspiration. The operational version of CREST model is
202 embedded in a modeling framework named Ensemble Framework for Flash Flooding Forecast (EF5,
203 https://github.com/HyDROSLab/EF5/releases) (Gourley et al., 2017) . The EF5 is an ensemble framework
204 with multiple hydrological model cores, including CREST version 2.0, co-developed by the University of
205 Oklahoma (http://hydro.ou.edu) and NASA Applied Science Team (www.servir.net) (Wang et al., 2011).

206 To build the CREST model in this study, we utilized the 30-arc-second (~1 km) hydrologically
207 conditioned Digital Elevation Model (DEM), Flow Direction (FDR), Flow Accumulation (FAA) data
208 obtained from the HydroSHEDS (https://www.hydrosheds.org/). For all four watersheds, the default
209 parameter set is based on a priori values developed by Vergara et al. (2016) for the CONUS at the same
210 spatial resolution (1 by 1 km) as the DEM data used in this study. By default, the automatic calibration was
211 done using the Differential Evolution Adaptive Metropolis (DREAM) algorithm (Vrugt et al., 2009)
212 available within the EF5 framework. The CREST model has 12 parameters as shown in Table 3 (Ma et al.,
213 2018).

214 (Figure 4 around here)

215 (Table 3 around here)

216 **2.2 Data-driven Machine Learning Models (DMLs)**

217 2.2.1 Artificial Neural Networks (ANN)

218 ANN model is a nonlinear model inspired by the working principle of human brains to mimic the
219 learning process (Jain et al., 1996; Leonard and Kramer, 1990). Human brains consist of a network
220 interconnected by billions of neurons. The neurons and synapses in the brain are expressed as nodes and
221 weights mathematically in the ANN model structure. The ANN model consists of interconnected networks
222 with layers, nodes, and weights called multi-layer fully-connected neural networks. Figure 5 shows (a) a
223 typical three-layer feed-forward ANN model and (b) the internal operations of the nodes in the hidden layer.
224 The n number of input data are assigned to the nodes in the input layer ($x_i, i = 1, 2, \dots, n$), and are
225 propagated through the interconnected nodes in the hidden layer with weight parameters and biases. Each
226 p number of hidden nodes receives the weighted sum of the input nodes ($z_j, j = 1, 2, \dots, p$), then they are
227 further transferred to the output layer through an activation function. (Rashid and Ahmad, 2016). It can be
228 described in Equation (1):

$$z_j = f_z\left(\sum_{i=1}^{n} x_i W_{ji} + c_j\right) \quad (1)$$

where z_j is the jth node in the hidden layer, W_{ji} is the weight parameter indicating the strength of the connections between the input and hidden nodes, c_j is the bias, and f_z is the activation function. Similarly, this process is also performed between the hidden layer and the output data in the output layer can be calculated by Equation (2):

$$y_m = f_y\left(\sum_{j=1}^{p} z_j W_{kj} + b_k\right) \quad (2)$$

where y_m is the output variable ($m = 1, 2, \ldots, k$), W_{kj} is the weight parameter indicating the strength of the connections between the hidden and output nodes, b_k is the bias, f_y is the activation function. This full connection enables the representation of nonlinear interactions between inputs and output (Chen and Billings, 1992; Zealand et al., 1999).

The key procedure of the ANN model is to find the best weight parameters combination using a training algorithm. The most common learning rule is the backpropagation (ASCE, 2000; Goh, 1995), which adjusts the weights of connections between neurons in hidden and output layers to reduce the error in the output during the training process. The optimal number of hidden nodes is tunable in designing the network, and they can be determined by trial-and-error or heuristic optimization algorithms (Valipour et al., 2013; Yang et al., 2017a). Besides, there are several types of activation functions available to process the information within a node, such as the sigmoid function, hyperbolic tangent function, and sign function (Hsu et al., 1995; Zealand et al., 1999).

(Figure 5 around here)

2.2.2 Long Short Term Memory (LSTM)

The LSTM model is a special kind of RNN designed to overcome the gradient vanishing and exploding problems in the RNN (Hochreiter and Schmidhuber, 1997; Kratzert et al., 2018; Rumelhart et

al., 1986; Rumelhart et al., 1994). The LSTM model often provides outstanding performance for time-series forecasts by connecting previous information to the present, specializing in predicting sequence data series. Figure 6(a) shows a folded LSTM model structure which is also known for a regular RNN model structure. The difference between the regular RNN and the LSTM model is the formulation of the LSTM memory cell, which is the recurrent connection on the hidden layer. This design allows the LSTM model to capture the long-term sequential information and deal with the temporal correlations in the input data. The LSTM model first initiates with the zero vectors of the cell state and hidden state and then updates those two states by passing them into an LSTM memory cell with the input data. Figure 6(b) shows the internal operations of an LSTM memory cell in each time step t. Two steps are taken inside an LSTM cell to update the cell state (c_{t-1}) and the hidden state (h_{t-1}) into c_t and h_t with the input data (x_t). The first step is generating three gates and one update vector according to Equations (3) to (6):

$$f_t = \sigma\big(W_f[X_t, h_{t-1}] + b_f\big) \tag{3}$$

$$i_t = \sigma(W_i[X_t, h_{t-1}] + b_i) \tag{4}$$

$$\tilde{c}_t = \tanh(W_c[X_t, h_{t-1}] + b_c) \tag{5}$$

$$o_t = \sigma(W_o[X_t, h_{t-1}] + b_o) \tag{6}$$

where σ is the sigmoid activation function with return values from 0 to 1, tanh is the hyperbolic tangent activation function with return values from -1 to 1. W_f, W_i, W_c, and W_o are weight parameters. b_f, b_i, b_c, and b_o are the bias, and f_t, i_t, $\tilde{c}_t$ and o_t are forget gate, input gate, update vector, and output gate, respectively. The second step is to update the cell state and the hidden state. Last, the cell state is updated by Equation (7), and the new hidden state can be updated by Equation (8):

$$c_t = f_t \times c_{t-1} + i_t \times \tilde{c}_t \tag{7}$$

$$h_t = o_t \times \tanh(c_t) \tag{8}$$

where × indicates pointwise multiplication, and the three gates f_t, i_t, o_t range from 0 to 1. The forget gate f_t controls the previous cell state (c_{t-1}) and $i_t \times \tilde{c}_t$ controls the current time step. The output data in the output layer (y) can be obtained by connecting LSTM cells and feeding the time series of n time steps into them with the zero-initialized cell state and hidden state (c_0 and h_0) as presented in Equation (9):

$$y = f(x_1, x_2, \dots, x_{n-1}, x_n, c_0, h_0) \tag{9}$$

(Figure 6 around here)

2.3 Statistical evaluation

Four statistical criteria are used to compare the model accuracy: The Root Mean Square Error (RMSE), Correlation Coefficient (CC), Kling-Gupta Efficiency (KGE) (Gupta et al., 2009), and Nash-Sutcliffe Efficiency (NSE) (Nash and Sutcliffe, 1970), as defined in Equations (10)-(13), respectively:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Q_{sim,i} - Q_{obs,i}\right)^2} \tag{10}$$

$$CC = \frac{\sum_{i=1}^{n}((Q_{sim,i}-\bar{Q}_{sim,i})(Q_{obs,i}-\bar{Q}_{obs,i}))}{\sqrt{\sum_{i=1}^{n}(Q_{sim,i}-\bar{Q}_{sim,i})^2 \sum_{i=1}^{n}(Q_{obs,i}-\bar{Q}_{obs,i})^2}} \tag{11}$$

$$NSE = 1 - \frac{\sum_{i=1}^{n}(Q_{obs,i}-\bar{Q}_{sim,i})^2}{\sum_{i=1}^{n}(Q_{obs,i}-\bar{Q}_{obs,i})^2} \tag{12}$$

$$KGE = 1 - \sqrt{(CC-1)^2 + (BR-1)^2 + (RV-1)^2} \tag{13}$$

where n is the data length, $Q_{obs,i}$ and $Q_{sim,i}$ are the observed and simulated data series, and $\bar{Q}_{obs,i}$ and $\bar{Q}_{sim,i}$ are the means of the observed and simulated data series. The BR and RV in the KGE can be calculated by Equations (14)-(15), respectively:

$$BR = \frac{\bar{Q}_{sim}}{\bar{Q}_{obs}} \tag{14}$$

$$RV = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Q_{sim,i}-\bar{Q}_{sim,i}\right)^2}/\bar{Q}_{sim}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Q_{obs,i}-\bar{Q}_{obs,i}\right)^2}/\bar{Q}_{obs}} \quad (15)$$

The RMSE measures the magnitude of the absolute error between the observed and simulated data. The CC is the linear correlation to identify either a negative or positive correlation between the observed and simulated data ranging from -1 to 1. The NSE is widely used to assess the model accuracy by calculating the errors between observed and simulated data and the variance of the observed data. The model accuracy can be evaluated as "very good" if $0.75<NSE\leq1$, "good" if $0.65<NSE\leq0.75$, "satisfactory" if $0.50<NSE\leq0.65$, and "unsatisfactory" if $NSE\leq0.50$ (Moriasi et al., 2007). The KGE is used to measure the model accuracy by identifying possible sources of systematic errors considering the three criteria such as the CC, the BR, and the RV. Both KGE and NSE range from -∞ to 1, and the ideal value is 1 (Baez-Villanueva et al., 2018; Gupta et al., 2009; Kling et al., 2012; Knoben et al., 2019; Waseem et al., 2017).

3. Study region and data

3.1 Target Basin

In this study, we selected four signature watersheds with different hydroclimatic conditions to compare the employed PHMs and DMLs. Table 4 presents the basic information of the four selected watersheds, and Figure 7 shows the locations of the selected four watersheds over the CONUS. More details are described as follows.

(Table 4 around here)

(Figure 7 around here)

The Bushkill watershed is located in the eastern part of Pennsylvania. The water from this watershed generally flows southeast directly into the Delaware River, and the Blue Mountains form the northern boundary of the basin (Germanoski, 1999). The upper part of the basin is dominated by forest and

agricultural and residential land uses. The lower part of the basin features karst carbonate terrain with low-relief hills (Brandes, 2001). Towards the southernmost area of the basin, the land use changes from agricultural to residential and urban (Ruggles et al., 2001). The climate over the Bushkill watershed area features rainfall generated by land-falling tropical cyclones, winter-spring extratropical systems, and warm-season convective systems (Smith et al., 2010). Floods in this area occur in response to the mixture of extreme rainfall forcing and land surface properties.

The Mill watershed is located in the northeastern part of Kansas. The landscape mainly consists of low hills covered by grassland and some trees in the floodplain, and the basin is underlain by aquifer systems recharged in Colorado near the Rocky Mountain Thrust Belt (York et al., 2002). The soil within the Mill generally consists of erosive to moderately erosive silt and silty clay loams (Evans, 2005). Most of the flow originates from overland or shallow subsurface since the percolation of precipitation to groundwater is largely limited due to impermeable limestone and shale bedrock (Lee et al., 2009).

The Wimberley watershed is located in the southcentral part of Texas. The basin is oriented in an east-southeast direction and forms the northern headwaters of the larger Guadalupe River system, which flows to the Gulf of Mexico (Furl et al., 2018). The main stem of Wimberley features surface water-groundwater interaction. The climate in Wimberley is classified as subtropical humid with short, mild winters and hot summers, and droughts can occur and persist for months or years (Nielsen-Gammon, 2011). Wet seasons are spring and fall, and most precipitation is caused by the passage of continental fronts, mesoscale convective systems, and localized convective events during warmer months (Lin et al., 2018; Smith et al., 2000). Occasionally, tropical rainfall systems protrude far enough inland resulting in extreme precipitation events (Asquith and Slade, 1995).

The Sycamore watershed is located in the southcentral part of Arizona. The basin originates from the Mazatzal Mountains and flows south-westwards to the Verde River, and it is mountainous with altitudes ranging from 420m to 2,160m above sea level (Thomsen and Schumann, 1969). The climate in Sycamore

has mild winters and dry, hot summers featuring high temperatures and low relative humidity which makes intense evapotranspiration causing intermittent flows (Fisher et al., 1982). The heavy precipitation in this area is usually from convective storms during the monsoon in August.

As shown in the previous studies, all four watersheds are signature watersheds and cover various dry/wet climate conditions without artificial controls. The annual mean precipitation of these selected watersheds ranges from 295 to 1150 mm (Table 4), and the geographical locations cover the western, central, and eastern US, covering elevation ranges from 340 to 1147m. The total basin area range is from 306 to 843km^2 (Figure 7 and Table 4).

3.2 Datasets

The hydrological simulation requires datasets such as precipitation, PET, and streamflow representing water cycle fluxes in hydrologic modeling. Three hydrologic datasets are used in this study: 1) daily precipitation from the Parameter-elevation Regressions on Independent Slopes Model (PRISM) climate datasets in the PRISM climate group at Oregon State University, 2) daily PET from Famine Early Warning Systems Network (FEWS NET) at the United States Geological Survey (USGS), and 3) daily streamflow from National Water Information System (NWIS) at the USGS. The data length used to simulate the hydrologic models is from Jan 01, 2002, to Dec 31, 2019 (total of 18 years).

The daily PRISM precipitation is widely used to obtain spatially distributed precipitation data based on a regression between elevation and observed precipitation (Ashfaq et al., 2016; Oubeidillah et al., 2014; Prat and Nelson, 2015; Widmann and Bretherton, 2000). The daily PRISM precipitation used in this study is gridded precipitation with a spatial resolution of 4km (Daly and Bryant, 2013). Detailed information on describing the PRISM precipitation is available from http://prism.oregonstate.edu. The FEWS NET daily PET is calculated from Global Data Assimilation System (GDAS) analysis fields from the 6-hourly numerical meteorological model output using the Penman-Monteith equation (Allen et al., 1998; Verdin et al., 2005). The daily PET used in this study is a gridded PET that has 1° by 1° resolution

(https://earlywarning.usgs.gov/fews). Note that the spatial resolutions of the PRISM and PET are bilinearly interpolated from their original to the 1km domain grids to deal with the different spatial resolutions between inputs. Daily streamflow data at the outlet of four selected watersheds are the quality-assured gauge data based on time-series format (http://waterdata.usgs.gov/usa/nwis/sw).

4. Model Setup

4.1 Input scenarios

In this study, we generated three input scenarios to examine the sensitivities of each model. Firstly, mean precipitation and mean PET at each grid are used as default inputs to drive both the PHMs and DMLs (the first input scenario; S1). Further, the second input scenario (S2) and the third input scenario (S3) were generated by adding delayed precipitation/PET to drive the DMLs only. The adding of delayed forcing data could guide the DMLs to capture how water is being delayed and routed through the watershed in the training process (Yang et al., 2019). To be specific, the S1 input scenario consists of mean precipitation and mean PET at each grid, and the lumped PHMs in this study (i.e., SAC-SMA and XAJ model) take the areal averaged mean value over the grids within each watershed. In the S2 input scenario, the lagged mean precipitation was added to the S1 input scenario. The lag time of the mean precipitation and mean PET was determined by the cross-correlation coefficient with the streamflow. For one-day lagged mean precipitation, significant cross-correlation was found with daily streamflow at all employed watersheds. However, for a one-day lagged mean PET, the significant cross-correlation between the mean PET and daily streamflow was found in the Bushkill and Sycamore watersheds. Therefore, the S2 input scenario in the Bushkill and Mill watersheds includes the S1 and one-day lagged mean precipitation and one-day lagged mean PET while the S2 in the Mill and Wimberley watersheds includes S1 and only one-day lagged mean precipitation. In the S3 input scenario, the lagged streamflow was further added along with the S2. The lag time of the streamflow was determined by the partial autocorrelation coefficient function (PACF). In Bushkill and Sycamore, one-, two-, and three- days were considered as additional model inputs because the spikes of the

PACF cut off at lag 3 in the streamflow autocorrelation experiment. In the Mill and Wimberley watersheds, only a one-day time step was added to the model inputs as the spike of the PACF mainly cut off at lag 1. In summary, the S1 was consistently applied for all employed PHMs and DMLs, while the S2 and S3 were individually set up when using the DMLs in order to examine whether DMLs could utilize flexible inputs and improve the streamflow simulation accuracy. These S1, S2, and S3 settings guarantee the comparison of model performance under rational and fair conditions. The detailed information for the three input scenarios are summarized in Table 5.

(Table 5 around here)

4.2 Parameters

Parameters play an important role in both PHMs and DMLs. The number of parameters of the SAC-SMA, XAJ, and CREST is respectively 16, 15, and 12(Tables 1-3). The parameters are optimized in the calibration set by tuning them within the defined ranges. Two lumped hydrologic models (SAC-SMA and XAJ models) are calibrated through the SC-SAHEL algorithm (Naeini et al., 2019; Rahnamay Naeini et al., 2018) with the objective function of minimum mean square error (MSE). We adopted the prior-calibrated CREST model parameters for these watersheds according to the study by Vergara et al. (2016). Vergara et al. (2016) used the DREAM algorithm and the spatial parameters across the US are available in the EF5 framework. We note that the PHMs parameter calibration could possibly bring uncertainties in streamflow simulation when using different calibration strategies. However, the main focus of this paper is to compare PHMs with DMLs instead of investigating the impacts of PHMs parameter calibration strategies. The use of the newly developed SC-SAHEL algorithm to calibrate all employed PHMs is possible, but it needs a re-building of the source code of the CREST model at a large scale, which will be explored in future studies. Nevertheless, both the DREAM algorithm (Vrugt et al., 2009) and SC-HAHEL algorithm (Naeini et al., 2019) are originated from the SCE-UA algorithm, which is the ancestor of these optimization algorithms (Duan et al., 1992), and we believe that the calibration performance will not be significantly

different. There are some structural differences between the DREAM algorithm and the SCE-UA algorithm. The DREAM algorithm has been known to exhibit excellent performance on model calibration studies as it provides a better estimation of the posterior distribution, while the SCE-UA algorithm still has limitations in solving premature convergence (Chu et al., 2014; Laloy and Vrugt, 2012; Naeini et al., 2019). Interested readers could conduct similar comparison studies with the CREST model and further explore whether calibration methods will dramatically change the model performance if switching from the DREAM algorithm to the SC-SAHEL algorithm.

In the DMLs, we set the number of hidden nodes in the ANN model and the number of hidden states in the LSTM model as hyperparameters. Thus, the ANN model considered a total of 20 hidden nodes from 1 to 20, and the LSTM model considered a total of 10 hidden states from 1 to 250 (1, 10, 20, 30, 40, 50, 100, 150, 200, 250). In common with ANN and LSTM models, all connections between nodes consist of weight parameters. The optimal weights are determined by tuning the networks starting with randomly generated initial parameters based on the objective function of minimum MSE. Simultaneously, the optimal number of hidden nodes (ANN model) and hidden states (LSTM model) are respectively selected through iterative experiments based on the minimum MSE. Apart from the hyperparameters, for the ANN model, the training algorithm was set to resilient backpropagation, and the activation function was set to the logistic sigmoid function. The number of epoch was set to 500, and the learning rate was set to 0.01. For the LSTM model, we set the training option to Adam optimizer. The cell and hidden states were updated using the hyperbolic tangent activation function, and gates were computed using the sigmoid activation function. The number of the epoch was set to 200, and the learning rate was set to 0.01.

4.3 Data splitting

Cross-Validation (CV) is the most commonly used data splitting method to evaluate model performance (Bergmeir and Benítez, 2012; Cawley and Talbot, 2010). It divides a given dataset into k number of sub-groups with an equal size of samples (k is a parameter, so-called k-fold). Each sub-group of

data becomes an independent validation set, and the remaining folds are used to train or calibrate a model (Xu and Goodacre, 2018). With this approach, the model could be validated total k times, using different sub-groups as calibration and validation. The final model performance can be obtained by averaging the model skills over all k-subsets, of which the k-th subset of data was used as validation independently. In hydrologic modeling, a similar concept is adopted to split the entire data into two subsets in temporal sequence, and the two subsets are called calibration and validation, respectively. While in machine learning and statistical model evaluation, the CV with 2+ folds are commonly accepted. The value of k could be determined by data availability, and generally from 2 to 10 (Zhou et al., 2017). This k-fold CV method helps prevent overfitting problems and avoid favoring some models to specific calibration and validation periods. In this study, we employed a three-fold CV (k=3) method with 12 years of data for calibration and 6 years of data for validation for both PHMs and DMLs (Figure 8). When separating the entire 12 years of data into 3 folds, the DMLs could be easily built and tested following common practice in evaluating machine learning techniques, in which the training, validation, and testing datasets were used (Wu et al., 2013). In addition, the PHMs could be validated a total of 3 times using different combinations of calibration and validation datasets. Note that for the three employed PHMs, the warm-up period was set from Feb 01, 2001, to Dec 31, 2001, for the model states to come to equilibrium.

(Figure 8 around here)

5. Results

Tables 6-9 show the model performance of simulating five employed models (SAC-SMA, CREST, XAJ, ANN, and LSTM) using k-fold CV in four selected watersheds. The statistical measurement is the average value of three folds (k=3) in each calibration and validation set. There is a total of nine cases since the PHMs (SAC-SMA, XAJ, and CREST) respectively have one result by the first input scenario (S1), and the DMLs (ANN and LSTM) respectively have three results by the three different input scenarios (S1, S2, and S3). In Tables 6-9, the bold-faced values are the best statistics, indicating the highest model accuracy

for each statistical criterion among the nine cases. According to the three input scenarios, the underlined values are the best statistics within the ANN and LSTM models. Based on the four statistical criteria comprehensively, the best performing ANN and LSTM models among different input scenarios is the ANN model under the S3 input scenario and the LSTM model under the S3 input scenario in the Bushkill watershed. In contrast, the best performing ANN and LSTM models in the rest three watersheds (Mill, Wimberley, and Sycamore) are the ANN model under the S3 input scenario and the LSTM model under the S1 input scenario. For the convenience of terminology, for instance, the ANN model under the first input scenarios is now referred to as ANN (S1), and the same applies to the remaining input scenarios.

In the Bushkill watershed (Table 6), the best RMSE, CC, NSE, and KGE values are obtained by the ANN (S3) in both calibration and validation sets. The best CC, NSE, and KGE values in the validation are respectively 0.96, 0.92, and 0.94, indicating outstanding streamflow simulation capability of the ANN (S3). Along with the ANN (S3), the LSTM (S3) shows the second-best streamflow simulation performance in the Bushkill watershed. In the Mill watershed (Table 7), the best RMSE, CC, NSE, and KGE values are obtained by the XAJ model in both calibration and validation sets. The best CC, NSE, and KGE values in the validation period are 0.83, 0.68, and 0.79, respectively, indicating moderately good statistical performance. Meanwhile, the ANN (S3) has better performance than the ANN (S1, S2), and the LSTM (S1) has better performance than the LSTM (S2, S3) by showing the best statistics within their categories. However, the performance of ANN (S3) is similar to that of the SAC-SMA model but worse than the XAJ model. The performance of LSTM (S1) is worse than that of ANN (S3). In the Wimberley watershed (Table 8), the best RMSE, CC, NSE, and KGE are alternately obtained by the SAC-SMA and the XAJ models in both calibration and validation sets. Especially for the XAJ model, the best CC, NSE, and KGE values in the validation set are 0.79, 0.51, and 0.47, respectively, indicating the satisfactory performance of the XAJ model in general. In the same Wimberley watershed, except for the XAJ model, all other models show unsatisfactory performance. For example, several calculated NSE and KGE values from other PHM and DMLs are close to zero or negative. In the Sycamore watershed (Table 9), the best RMSE, NSE, and KGE

481 values are obtained by the ANN (S3) model, and the SAC-SMA model obtains the best CC over the
482 validation set. However, the NSE and KGE values obtained by the SAC-SMA model are much lower than
483 those from the ANN (S3) model. Besides, the LSTM (S1) model shows the best statistics, and the NSE
484 value from its obtained simulation result is better than the scores achieved by other PHMs during the
485 validation set. Overall, the performance of PHMs and DMLs varies considerably for each selected
486 watershed, but several models obtain significantly satisfactory daily streamflow simulations under different
487 input scenarios.

488 (Tables 6-9 around here)

489 Figures 9-12 show the time series plots of simulation results for each k-fold (k=3) in the four studied
490 watersheds. In the figures, there are three subsets according to the 3-fold CV, and each fold has its
491 calibration and validation sets. We also present the simulation results divided into (a) PHMs and (b) DMLs
492 for easier reading and understanding. In Figures 9-12, the black line represents the observed daily
493 streamflow. In Figures 9(a)-12(a), the blue, orange-, green- solid lines are the simulation results from the
494 SAC-SMA, XAJ, and CREST models, respectively. In Figures 9(b)-12(b), the red-, yellow- solid lines are
495 the simulation results of ANN and LSTM models, respectively. The simulation results of ANN and LSTM
496 models inhere represented by the results with the best performing input scenarios (S1, S3, S3). Therefore,
497 the time series plots shown as DMLs are ANN (S3) and LSTM (S3) for the Bushkill watershed, and are
498 ANN (S3) and LSTM (S1) for the Mill, Wimberley, and Sycamore watersheds. In the Bushkill and Mill
499 watersheds (Figures 9-10), throughout the calibration and validation set at all folds, all models match well
500 the overall patterns of the observed streamflow. In the Wimberley and Sycamore watersheds (Figure 11-
501 12), the number of peak flows has much less than the two previous watersheds. In general, the SAC-SMA,
502 XAJ, and ANN models are able to match the peak flows than other models better, while the CREST model
503 shows some overestimations in all flow regimes. As shown in those figures, the observed streamflow in the
504 four watersheds have different temporal properties, particularly on a seasonal basis, and the model

505 performance varies accordingly. Therefore, additional exploration considering seasonality is needed. In the
506 following section, we further investigate the seasonal performances of each model.

507 (Figures 9-12 around here)

508 Figure 13 shows the scatter plots between observed and simulated streamflow during validation set
509 on a seasonal basis for the four watersheds. For better clarity, the results are presented in logarithm axes.
510 There are plots of the Bushkill, Mill, Wimberley, and Sycamore watersheds from top to bottom, and the
511 month of DJF, MAM, JJA, and SON are placed from left to right (i.e., DJF, MAM, JJA, and SON refers to
512 December-January-February, March-April-May, June-July-August, and September-October-November,
513 respectively). The zeros in both observation and simulation are ignored in this figure. The 1:1 diagonal line
514 indicates a perfect simulation that all simulated values match the observation, and the closer the plotted
515 points to the 1:1 line, the more reliable the simulations. In the Bushkill watershed (the first row in Figure
516 13), the data points from the five models generally lined up along the 1:1 line. Among the four seasons, the
517 DJF and MAM results show better matches to observation than those during the months of JJA and SON.
518 Further, the ANN model shows the most reliable simulations in general as they are well-matched to the 1:1
519 line. Other models also show well-matched to the 1:1 line; however, the scatter plots results tend to be
520 slightly dispersed. In the Mill watershed (the second row in Figure 13), the scatter plots from all PHMs and
521 DMLs are more dispersed than the Bushkill watershed but centered on the 1:1 line. There are prominent
522 commons throughout four seasons that the simulations from the ANN models are above the 1:1 line in the
523 low-flow regime, while the SAC-SMA model simulations are below the 1:1 line. The XAJ model
524 simulations tend to show dispersion in DJF and MAM, but still follow the 1:1 line. However, the XAJ
525 model's results are mostly above the 1:1 line in JJA and SON. In the Wimberley watershed (the third row
526 in Figure 13), the data points from all PHMs and DMLs are fairly dispersed. Particularly, the ANN model
527 simulations appear to be the poorest in matching the observations throughout all seasons, while the XAJ
528 model's results show better performance in MAM and JJA than that in DJF and SON. In the Sycamore

529 watershed (the last row in Figure 13), there are very few plotted data points in JJA and SON, indicating dry
530 conditions during these seasons. The ANN model simulations are well-matched to the 1:1 line than the
531 other models. Based on the seasonal scale analysis, the results from both the PHMs and DMLs vary greatly
532 depending on the watershed behaviors across different seasons and the magnitudes of flows in each
533 watershed. Accordingly, a detailed comparison of model performance related to the climate condition of
534 the watersheds will be provided in later discussion section. Based on the findings at the seasonal scale
535 presented here, we will further investigate the model behaviors concerning the magnitude of flows at each
536 watershed.

537 (Figure 13 around here)

538 To examine how the PHMs and DMLs will perform under different magnitudes of flow regimes,
539 we calculate the biases between simulated and observed flows, specifically for three different flow regimes:
540 the high-flow regime (≥80%), mid-flow regime (20%<and <80%), and low-flow regime (≤20%) (Fang
541 and Shen, 2017), for each studied watershed. Note that the zero values in the observation are also ignored
542 when separating flow regimes, but the zero values are included when calculating the model biases. In this
543 study, we use the mean absolute error (MAE) to represent model biases. Table 10 shows the calculated
544 MAE in different flow regimes (High-flow, Mid-flow, and Low-flow) during the validation set for each
545 model for the four studied watersheds. The bolded and underlined values represent the best statistics among
546 all PHMs and DMLs under each flow regime.

547 In the high-flow regime, three watersheds (i.e., the Bushkill, Mill, and Sycamore) consistently show
548 the lowest MAE when simulating the streamflow using the ANN model, except for the Wimberley
549 watershed where the LSTM model obtains the lowest MAE. This indicates that the DMLs have
550 considerably more accurate model performance in reducing the simulation biases in the high-flow regimes.
551 In the mid-flow regime, two watersheds show the lowest MAE with the ANN model, and the other two
552 watersheds show the lowest MAE with the results obtained by the SAC-SMA model. However, in the low-

553 flow regime, the lowest MAEs are mostly obtained by the SAC-SMA model except for the Bushkill
554 watershed, where the ANN model obtains the lowest MAE. This result suggests that in a low-flow regime,
555 the PHMs have better performance in terms of biases than the DML models. While in the medium- to high-
556 flow regimes, the DMLs are more capable than PHMs in reducing the simulation biases.

557 (Table 10 around here)

558 **6. Discussion**

559 This study employed three PHMs and two DMLs to cross-evaluate their capabilities in simulating
560 daily streamflow for four watersheds with different climate and geomorphological conditions. We
561 specifically investigated the general statistics of streamflow simulation in 3-fold cross-validation, the
562 model's performance over different seasons, and the simulation accuracy over high-, medium- and low-
563 flow regimes. For all employed models, mean precipitation and mean PET at each grid were used as default
564 input scenario (S1). Additional input scenarios (S2 and S3) were considered by adding delayed
565 precipitation/PET for the DMLs to enable the DMLs to capture the correlation in sequential data (Yang et
566 al., 2019). The k-fold CV (k=3) prevented models from favoring specific calibration or validation sets, and
567 the final simulation results were obtained by averaging the results from three subsets. We used the four
568 statistical criteria (RMSE, CC, NSE, and KGE) to assess and compare the model performances.

569 A pure comparison between PHMs and DMLs should come under the baseline input scenario (S1).
570 Tables 6-9 show that the ANN (S1) and LSTM (S1) generally showed similar or worse performance than
571 the SAC-SMA and XAJ models but slightly better performance than the CREST model. To be specific, the
572 CC values of the SAC-SMA, XAJ, CREST, ANN (S1), and LSTM (S1) in the validation set at the four
573 watersheds were around or above 0.5, indicating the simulated streamflow by all employed baseline models
574 matches well with general patterns in observations. However, the RMSE, NSE, and KGE values over the
575 validation set at all four watersheds showed significant differences among the employed PHMs and DMLs.
576 This result indicates that when using the same input information, the performance of DMLs is comparable

577 in time correlations to the PHMs. Still, the simulated results may be significantly biased when looking at
578 RMSE, KGE, and NSE. However, the DMLs are designed to have the flexibility to take multiple inputs,
579 not limited to the baseline input scenarios. This enables the use of delayed precipitation, PET, and
580 streamflow time series for DMLs to further capture the historical relationship between inputs and
581 streamflow. After adding delayed forcing data, it showed that the DMLs were able to derive better
582 streamflow simulation under S2 and S3 than the under baseline scenario S1. Besides, it is noteworthy that
583 the ANN model showed higher accuracy than the PHMs when given additional inputs in several cases. The
584 comparison between the PHMs and the best performing DMLs under different input scenarios demonstrates
585 that the DMLs have a great potential and can reach high simulation accuracy over the validation set. When
586 examining the performance among the employed PHMs, the XAJ model showed a lower NSE value than
587 the SAC-SMA model. This agrees with the previous studies by Xiang et al. (2016) and Huo et al., (2019),
588 as they showed the XAJ model generally had better NSE values than the SAC-SMA model. Meanwhile,
589 the CREST model showed more unsatisfactory performance than the two lumped models in our studied
590 watersheds. This also agrees with the previous DMIP1&2 studies by Smith et al. (2004), as one of the main
591 conclusions was that lumped and distributed models have no significant differences in their performance
592 although the distributed hydrologic models are newer developments over the classical lumped models.

593 Among all DMLs, our results show that the ANN model generally showed better performance than
594 the LSTM model. According to our results, the ANN model could still be a strong and robust model in
595 simulating the daily streamflow unlike previous results that showed the LSTM model generally performed
596 better than the ANN on average (Srivastava and Lessmann, 2018). Thus, this result implies that even with
597 more advanced DMLs, the classic DMLs can still perform comparably or even better than the deep learning
598 algorithms if proper inputs are given in the model training process. The Deep learning algorithm, such as
599 LSTM model, may work better in other complex datasets with a high level of randomness and noises, as
600 shown in many works in literature, but not the case in the hydrologic simulation when the river discharge
601 was governed by the physical routing process of precipitation and PET.

602 To further compare the PHMs and DMLs, we examined the simulation results over the validation
603 set, separately on the three different flow regimes (Table 10). Our findings showed that the DMLs had
604 better performance and a lower MAE value in the high-flow regime, while the PHMs had a better
605 performance for the low-flow regime. In the mid-flow regime, it is unclear whether PHMs had a better
606 performance for the low-flow regime. In hydrology, managing water in a high-flow regime is very
607 important to mitigate the impacts and reduce the risk of economic losses due to potential floods. A low-
608 flow regime reflects the continuous supply of water to agriculture sectors or groundwater recharge. From
609 this perspective, both PHMs and DMLs have their own strengths. Specifically, among the employed PHMs
610 and DMLs, in the Mill watershed, the ANN model showed the lowest MAE values in both high- and mid-
611 flow regimes, and the SAC-SMA model showed the lowest MAE value in the low-flow regime. However,
612 the previous result showed that when considering all flow regimes, the best statistics in the Mill watershed
613 were obtained by the XAJ model (Table 7). This result implies that the PHMs and the DMLs have different
614 performances, and the model performances depend on the flow regimes. It also suggests that one single
615 model may perform the best over an entire and continuous validation set but cannot consistently prevail in
616 a particular flow regime. Based on this result, we can infer that the multi-model ensemble of both PHMs
617 and DMLs over different flow regimes may generate a better simulation than any single hydrologic model
618 over the validation set, which is another research topic beyond the scope of work of this study.

619 Based on the results, we found seasonality plays an important role in regional hydrology. The
620 performances of different types of hydrologic models, either PHMs and DMLs, vary based on the seasonal
621 variations of streamflow and precipitation. Table 11 presents the mean and coefficient of variation (CV) of
622 seasonal streamflow at the four watersheds. The streamflow in the Bushkill watershed had a relatively
623 strong seasonal variation, with the largest streamflow in MAM than other seasons. The CV values during
624 the four seasons in the Bushkill watershed were all close to 1, indicating the streamflow demonstrates a
625 similar pattern with precipitation, and the rainfall-runoff relationship is very consistent over the entire
626 period. This tells us two things. First, both the PHM and the DMLs are able to provide satisfactory

627 simulation accuracy under stable climate conditions, where streamflow closely relates to the seasonal
628 variation in precipitation. Second, however, the ANN model is able to derive the highest statistics values
629 among all employed models, only if the model can fully use the historical observations with the delayed
630 information. The streamflow in the Mill watershed also had the strongest rainfall-runoff relationship than
631 the other three watersheds. The CV values in the Mill watershed were about three times larger than that in
632 the Bushkill watershed; however, there were no significant differences among different seasons. The high
633 simulation accuracy and statistic values obtained in the Mill watershed also indicate that the PHMs and the
634 DMLs are interchangeable across different seasons. The streamflow in the Wimberley watershed is featured
635 by strong surface-groundwater interaction; therefore, the physical processes under the surface are the main
636 contributors to total runoff. However, the CV values in the Wimberley watershed were considerably
637 different in various seasons, with a relatively high CV value recorded in JJA. The overall CV values
638 obtained by all models in the Wimberley watershed were poorer than the results from other watersheds. In
639 many modeling cases for the Wimberley watershed, both PHMs and DMLs showed some levels of
640 overfitting, except for the XAJ model. In this Wimberley watershed, the XAJ model showed a better
641 performance than any DMLs. The other two PHMs also showed similar or even worse statistical
642 performances than the DMLs. The streamflow in the Sycamore watershed is dominated by an intense
643 evaporation process, which caused low and intermittent flows. Notably, the CV value in JJA was very high
644 in the Sycamore watershed due to the winter convective storms, which bring a number of heavy
645 precipitation events. The high CC values in JJA indicate both PHMs and DMLs could capture the rainfall-
646 runoff relationship in the Sycamore watershed when a large amount of water falling on the watershed.
647 However, the process is different between PHMs and DMLs. The PHMs use the physical water routing
648 process and water balance to transform the heavy precipitation into runoffs. The DMLs try to establish the
649 statistical relationship between precipitation and runoff. From the training process, the DMLs will identify
650 that whenever there is heavy precipitation occurs, a significant increase in the streamflow time series is

651 followed. This is also why the scatter plots result in the ANN model demonstrates the lowest MAE in the
652 Sycamore watershed's high-flow regime, of which most of the high flows occur in wintertime.

653 In a nutshell, the PHMs are limited in capturing flow regimes in some seasons and flow regimes.
654 This is because the PHM has a basic assumption with fixed model structure, input, and parameter definition
655 to describe the physical process. Model parameter calibration could help with the PHMs to better capture
656 the rainfall-runoff process; however, there is a limitation on how much improvement the calibration can
657 achieve in a fixed model framework of all types of PHMs. While DMLs are free of assumptions, the DMLs
658 are more flexible to capture the dynamics than the PHMs in using different input scenarios and even use
659 climate, season, and other representative predictors to help establish the rainfall-runoff relationship through
660 training. The limitation of DMLs is the lack of a mass-balance check on how much water being transformed
661 from precipitation and how much water remains within the watershed. The authors believe that both
662 modeling frameworks can be further developed to benefit each other.

663 Last but not least, any comparison of PHMs and DMLs will be somehow biased, and there is no
664 fully fair foundation. That's because the PHMs are well-established tools, and they have been continuously
665 improved by hydrologists, engineers, and researchers for over 40 years. The PHMs are currently supported
666 by many existing studies, experienced users, government supports, and a health research and operation
667 environment for further development. This is a foundation that the DMLs do not have yet, as they are the
668 newer tools, especially with respect to hydrological science applications. From this perspective, any
669 comparison study is usually biased in favor of the established framework, the PHMs. Beyond the models'
670 transferability we discussed above, the methods for regionalization, i.e., to transfer data from similar donor
671 catchments to the catchment under consideration, is such a resource that classic PHMs can use but DMLs
672 are still lacking. However, the transfer of such data is, in principle, a big asset for DMLs by using similarity
673 measures (Kratzert et al. 2019b). With more and more researchers and young generations of hydrologists
674 investigating the usefulness of DMLs in hydrological modeling, we expect the research environment and

user foundation gap between PHMs and DMLs will be further closed. For example, researchers have used the concepts of connecting different water buckets to represent water flows and storage in different soil layers in the PHMs' framework. Similar concepts could also be implemented with DMLs' framework by connecting a few DMLs separately trained for each soil layer. Classical PHMs have already drawn on those existing understanding to develop further. As our understanding of the difference between PHMs and DMLs improves, our knowledge of the existing water cycle and the limitations of PHMs could quickly help us improve the DMLs' performance, and authors remain a positive attitude about the future of DMLs in hydrological modeling.

7. Conclusion

In this study, we figured out the streamflow simulation capability of PHMs and DMLs over four selected watersheds in the CONUS. To do this, three PHMs (SAC-SMA, XAJ, and CREST), and two DMLs (ANN and LSTM) were tested under different input scenarios. We first evaluated the overall simulated results against observed streamflow using four statistics, and investigated whether PHMs and DMLs have varying performances over different seasons and flow regimes. Our findings provided an in-depth examination of the functional difference between PHMs and DMLs and their simulation capability and accuracy under different flow regimes, climate, and geomorphological conditions.

Based on the simulation experiment results, we concluded that DMLs are quite competitive in many cases compared to the PHMs. For example, in the Bushkill watershed, the DMLs fed with delayed input scenarios (S3) demonstrate superior statistical performances over all the PHMs. At the same time, it should also be noted that the ANN provides the best accuracy only when sufficient information is provided as inputs. This is because there is no physical water routing process being formulated in the DMLs. When the inputs to DMLs did not directly carry the rainfall-runoff relationship, the runoff always responds by capturing dynamics in a delayed manner. This also could appear as a weakness for all DMLs because the DMLs are extremely sensitive to the inputs (Kim et al., 2020). Whenever the inputs do not have such a

delayed relationship between feature variables and target variables, all DMLs tend to fail at identifying the watershed runoff generation process. However, the DMLs are designed to flex in taking multiple and various input variables. Therefore, this limitation of DMLs could be manually offset by creating delayed precipitation, PET, and runoff as additional inputs. As a result, in this study, DMLs are promising tools and even superior in simulating the daily streamflow than the PHMs in certain regions, flow regimes, and seasons under proper model input scenarios.

We also found the DMLs have shown consistent better performance in the high-flow regime (≥80% of all data samples) compared to the baseline PHMs for all employed watersheds. This finding indicates that the DMLs are suitable tools for practical users want to forecast potential floods for risk mitigation. On the other hand, the PHMs maintain the best statistic in the low-flow regime and are more reliable tools for water resources planning in agriculture sectors and groundwater management. This finding differs from the previous studies, as typical lumped models often have problems representing low-flow regimes due to poor groundwater storage representation (Davison and van der Kamp, 2008). In the present study, the DMLs tend to overestimate the daily streamflow in the low-flow regime. This result reasonably implies that both PHMs and DMLs have their own merits and are worthy of joint development. Future studies may include developing hybrid forms of hydrologic models or a multi-model ensemble of PHMs and DMLs to simulate the streamflow under varying climate and different flow regimes.

In summary, leveraging Machine Learning algorithms to assist hydrologic simulation is becoming more and more popular in the field of hydrologic science. We need to carefully examine new developments and do not oversell the AI technologies in the fields of hydrology and water resources management. Systematical and large-scale evaluation studies are still required to fully vet the pros and cons when using new AI technologies in hydrological modeling and other topics. Thus, we suggest additional testing and comparison among different types of the PHMs and the DMLs to fully evaluate the usefulness of the DMLs under different climate and geomorphological conditions. More applications and comparisons should be

continued by adding more complex and miscellaneous climate and geomorphological conditions, such as snow-dominated watersheds and mountainous terrain, to further evaluate the streamflow simulation capabilities of both the PHMs and the DMLs. Large-scale comparison studies are strongly encouraged. Similar to those DMIP 1&2 studies initiated by the National Weather Service, the large-scale comparison between PHMs and DMLs shall be initiated by government agencies or relevant research centers to enrich our understanding of the different modeling approaches. We foresee the future development of hydrologic models toward the hybrid form of the DMLs and the PHMs to take advantage of the pros of both types of models while overcoming the cons of individual models and associated data requirements.

8. Data Availability

Daily PRISM precipitation, daily FEWS NET PET, and daily streamflow data used in this study are respectively available on the PRISM Climate Group (http://prism.oregonstate.edu), the USGS FEWS NET Data Portal (https://earlywarning.usgs.gov/fews), and the USGS Surface-Water Data for USA (http://waterdata.usgs.gov/usa/nwis/sw).

The original SAC-SMA model code is written in Fortran and is publicly accessible. This study used the MATLAB version translated from the Fortran code with the same version applied by Chu et al. (2011a; 2011b). The XAJ model is accessible at https://github.com/Ding-Ziyu/Xinanjiang-hydrological-model-for-MATLAB. The CREST model is available at https://github.com/HyDROSLab/EF5/releases. The ANNs and LSTM models were built using MATLAB R2019b software.

9. Acknowledgment

This work is partially supported by the U.S. Department of Energy (DOE Prime Award # DE-IA0000018). The material is based upon work supported by the National Science Foundation under Grant No. OIA-1946093 and its subaward No. EPSCoR-2020-3, and the National Science Foundation under Grant No. NSF1802872. The financial support is also made available by the National Key R&D Program of China under Grant No. 2018YFC0407902.

References

Abbot, J. and Marohasy, J. 2012. Application of artificial neural networks to rainfall forecasting in Queensland, Australia. Advances in Atmospheric Sciences 29(4), 717-730.

Abdulla, F., Lettenmaier, D. and Liang, X. 1999. Estimation of the ARNO model baseflow parameters using daily streamflow data. Journal of Hydrology 222(1-4), 37-54.

Ajami, N.K., Gupta, H., Wagener, T. and Sorooshian, S. 2004. Calibration of a semi-distributed hydrologic model for streamflow estimation along a river system. Journal of hydrology 298(1-4), 112-135.

Akbari Asanjan, A., Yang, T., Hsu, K., Sorooshian, S., Lin, J. and Peng, Q. 2018. Short-Term Precipitation Forecast Based on the PERSIANN System and LSTM Recurrent Neural Networks. Journal of Geophysical Research: Atmospheres 123(22), 12,543-512,563.

Aksoy, H. and Dahamsheh, A. 2009. Artificial neural network models for forecasting monthly precipitation in Jordan. Stochastic Environmental Research and Risk Assessment 23(7), 917-931.

Allen, R.G., Pereira, L.S., Raes, D. and Smith, M. 1998. Crop evapotranspiration-Guidelines for computing crop water requirements-FAO Irrigation and drainage paper 56. Fao, Rome 300(9), D05109.

ASCE 2000. Artificial neural networks in hydrology. I: Preliminary concepts. Journal of Hydrologic Engineering 5(2), 115-123.

Ashfaq, M., Rastogi, D., Mei, R., Kao, S.C., Gangrade, S., Naz, B.S. and Touma, D. 2016. High-resolution ensemble projections of near-term regional climate over the continental United States. Journal of Geophysical Research: Atmospheres 121(17), 9943-9963.

Asquith, W.H. and Slade, R.M. (1995) Documented and potential extreme peak discharges and relation between potential extreme peak discharges and probable maximum flood peak discharges in Texas, US Department of the Interior, US Geological Survey.

Aytek, A., Asce, M. and Alp, M. 2008. An application of artificial intelligence for rainfall-runoff modeling. Journal of Earth System Science 117(2), 145-155.

Azadi, S. and Sepaskhah, A.R. 2012. Annual precipitation forecast for west, southwest, and south provinces of Iran using artificial neural networks. Theoretical and Applied Climatology 109(1), 175-189.

Baez-Villanueva, O.M., Zambrano-Bigiarini, M., Ribbe, L., Nauditt, A., Giraldo-Osorio, J.D. and Thinh, N.X. 2018. Temporal and spatial evaluation of satellite rainfall estimates over different regions in Latin-America. Atmospheric Research 213, 34-50.

Behrangi, A., Khakbaz, B., Jaw, T.C., AghaKouchak, A., Hsu, K. and Sorooshian, S. 2011. Hydrologic evaluation of satellite precipitation products over a mid-size basin. Journal of Hydrology 397(3), 225-237.

Bergmeir, C. and Benítez, J.M. 2012. On the use of cross-validation for time series predictor evaluation. Information Sciences 191, 192-213.

Boyle, D.P., Gupta, H.V., Sorooshian, S., Koren, V., Zhang, Z. and Smith, M. 2001. Toward improved streamflow forecasts: value of semidistributed modeling. Water Resources Research 37(11), 2749-2759.

Brandes, D. (2001) Urban Drainage Modeling, pp. 808-817.

Brazil, L. (1989) Multilevel calibration strategy for complex hydrologic simulation models, US Department of Commerce, National Oceanic and Atmospheric Administration ….

Brown, M.E., Racoviteanu, A.E., Tarboton, D.G., Gupta, A.S., Nigro, J., Policelli, F., Habib, S., Tokay, M., Shrestha, M.S., Bajracharya, S., Hummel, P., Gray, M., Duda, P., Zaitchik, B., Mahat, V., Artan, G. and Tokar, S. 2014. An integrated modeling system for estimating glacier and snow melt driven

streamflow from remote sensing and earth system data products in the Himalayas. Journal of Hydrology 519, 1859-1869.
Buchtele, J., Elias, V., Tesar, M. and Herrmann, A. 1996. Runoff components simulated by rainfallrunoff models. Hydrological sciences journal 41(1), 49-60.
Burnash, R.J., Ferral, R.L. and McGuire, R.A. (1973) A generalized streamflow simulation system: Conceptual modeling for digital computers, US Department of Commerce, National Weather Service, and State of California ….
Burnash, R.J.C.m.o.w.h. 1995. The NWS river forecast system-catchment modeling. 311-366.
Cawley, G.C. and Talbot, N.L. 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. The Journal of Machine Learning Research 11, 2079-2107.
Chaney, N.W., Huijgevoort, M.H.J.V., Shevliakova, E., Malyshev, S., Milly, P.C.D., Gauthier, P.P.G. and Sulman, B.N. 2018. Harnessing big data to rethink land heterogeneity in Earth system models. Hydrology and Earth System Sciences 22(6), 3311-3330.
Chaney, N.W., Wood, E.F., McBratney, A.B., Hempel, J.W., Nauman, T.W., Brungard, C.W. and Odgers, N.P. 2016. POLARIS: A 30-meter probabilistic soil series map of the contiguous United States. Geoderma 274, 54–67.
Chen, S. and Billings, S. 1992. Neural networks for nonlinear dynamic system modelling and identification. International journal of control 56(2), 319-346.
Chow, V.T. 1964. Handbook of applied hydrology.
Chu, W., Gao, X. and Sorooshian, S. 2010. Improving the shuffled complex evolution scheme for optimization of complex nonlinear hydrological systems: Application to the calibration of the Sacramento soil-moisture accounting model. Water Resources Research 46(9).
Chu, W., Gao, X. and Sorooshian, S. 2011a. A new evolutionary search strategy for global optimization of high-dimensional problems. Information Sciences 181(22), 4909-4927.
Chu, W., Gao, X. and Sorooshian, S. 2011b. A solution to the crucial problem of population degeneration in high-dimensional evolutionary optimization. IEEE Systems Journal 5(3), 362-373.
Chu, W., Yang, T. and Gao, X. 2014. Comment on “High-dimensional posterior exploration of hydrologic models using multiple-try DREAM (ZS) and high-performance computing” by Eric Laloy and Jasper A. Vrugt. Water Resources Research 50(3), 2775-2780.
Clark, M.P., Nijssen, B., Lundquist, J.D., Kavetski, D., Rupp, D.E., Woods, R.A., Freer, J.E., Gutmann, E.D., Wood, A.W., Brekke, L.D., Arnold, J.R., Gochis, D.J. and Rasmussen, R.M. 2015. A unified approach for process-based hydrologic modeling: 1. Modeling concept. Water Resources Research 51(4), 2498-2514.
Dahamsheh, A. and Aksoy, H. 2009. Artificial neural network models for forecasting intermittent monthly precipitation in arid regions. Meteorological Applications 16(3), 325-337.
Daliakopoulos, I.N. and Tsanis, I.K. 2016. Comparison of an artificial neural network and a conceptual rainfall–runoff model in the simulation of ephemeral streamflow. Hydrological Sciences Journal 61(15), 2763-2774.
Daly, C. and Bryant, K. 2013. The PRISM climate and weather system—an introduction. Corvallis, OR: PRISM climate group.
Davison, B. and van der Kamp, G. 2008. Low-Flows in Deterministic Modelling: A Brief Review. Canadian Water Resources Journal / Revue canadienne des ressources hydriques 33(2), 181-194.
Duan, Q., Sorooshian, S. and Gupta, V. 1992. Effective and efficient global optimization for conceptual rainfall-runoff models. Water resources research 28(4), 1015-1031.
Evans, B.C. (2005) Soil Survey of Johnson County, Kansas, United States Department of Agriculture, Natural Resources Conservation Service.

Fang, K. and Shen, C. 2017. Full-flow-regime storage-streamflow correlation patterns provide insights into hydrologic functioning over the continental US. Water Resources Research 53(9), 8064-8083.
Fisher, S.G., Gray, L.J., Grimm, N.B. and Busch, D.E. 1982. Temporal succession in a desert stream ecosystem following flash flooding. Ecological monographs 52(1), 93-110.
French, M.N., Krajewski, W.F. and Cuykendall, R.R. 1992. Rainfall forecasting in space and time using a neural network. Journal of Hydrology 137(1), 1-31.
Furl, C., Sharif, H., Zeitler, J.W., El Hassan, A. and Joseph, J. 2018. Hydrometeorology of the catastrophic Blanco river flood in South Texas, May 2015. Journal of Hydrology: Regional Studies 15, 90-104.
Gan, Y., Duan, Q., Gong, W., Tong, C., Sun, Y., Chu, W., Ye, A., Miao, C. and Di, Z. 2014. A comprehensive evaluation of various sensitivity analysis methods: A case study with a hydrological model. Environmental Modelling & Software 51, 269-285.
Germanoski, D. 1999 The Lehigh Valley landform assemblage: differential erosion and relationships between topography and geology, pp. 9-30.
Goh, A.T. 1995. Back-propagation neural networks for modeling complex systems. Artificial Intelligence in Engineering 9(3), 143-151.
Gourley, J.J., Flamig, Z.L., Vergara, H., Kirstetter, P.-E., Clark III, R.A., Argyle, E., Arthur, A., Martinaitis, S., Terti, G. and Erlingis, J.M. 2017. The FLASH Project: improving the tools for flash flood monitoring and prediction across the united states. Bulletin of the American Meteorological Society 98(2), 361-372.
Grayson, R.B., Moore, I.D. and McMahon, T.A. 1992. Physically based hydrologic modeling: 1. A terrain-based model for investigative purposes. Water Resources Research 28(10), 2639-2658.
Gupta, H.V., Kling, H., Yilmaz, K.K. and Martinez, G.F. 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. Journal of Hydrology 377(1), 80-91.
Ham, Y.-G., Kim, J.-H. and Luo, J.-J. 2019. Deep learning for multi-year ENSO forecasts. Nature 573(7775), 568-572.
Hao, G., Li, J., Song, L., Li, H. and Li, Z. 2018. Comparison between the TOPMODEL and the Xin'anjiang model and their application to rainfall runoff simulation in semi-humid regions. Environmental Earth Sciences 77(7), 279.
Hochreiter, S. and Schmidhuber, J. 1997. Long short-term memory. Neural computation 9(8), 1735-1780.
Hogue, T.S., Gupta, H. and Sorooshian, S. 2006. A 'user-friendly'approach to parameter estimation in hydrologic models. Journal of Hydrology 320(1-2), 202-217.
Hogue, T.S., Sorooshian, S., Gupta, H., Holz, A. and Braatz, D. 2000. A multistep automatic calibration scheme for river forecasting models. Journal of Hydrometeorology 1(6), 524-542.
Hsu, K.l., Gupta, H.V. and Sorooshian, S. 1995. Artificial neural network modeling of the rainfall-runoff process. Water resources research 31(10), 2517-2530.
Hu, C.H., Guo, S.L., Xiong, L.H. and Peng, D.Z. 2005. A modified Xinanjiang model and its application in northern China. Nord Hydrol 36(2), 175-192.
Huang, P., Li, Z., Chen, J., Li, Q. and Yao, C. 2016. Event-based hydrological modeling for detecting dominant hydrological process and suitable model strategy for semi-arid catchments. Journal of Hydrology 542, 292-303.
Hung, N.Q., Babel, M.S., Weesakul, S. and Tripathi, N.K. 2009. An artificial neural network model for rainfall forecasting in Bangkok, Thailand. Hydrol. Earth Syst. Sci. 13(8), 1413-1425.

884 Huo, W., Li, Z., Wang, J., Yao, C., Zhang, K. and Huang, Y. 2019. Multiple hydrological models
885 comparison and an improved Bayesian model averaging approach for ensemble prediction over
886 semi-humid regions. Stochastic Environmental Research and Risk Assessment 33(1), 217-238.
887 Islam, Z. 2011. A review on physically based hydrologic modeling. University of Alberta: Edmonton, AB,
888 Canada.
889 Jain, A.K., Mao, J. and Mohiuddin, K.M. 1996. Artificial neural networks: A tutorial. Computer 29(3), 31-
890 44.
891 Ju, Q., Yu, Z., Hao, Z., Ou, G., Zhao, J. and Liu, D. 2009. Division-based rainfall-runoff simulations with BP
892 neural networks and Xinanjiang model. Neurocomputing 72(13-15), 2873-2883.
893 Kan, G., Tang, G., Yang, Y., Hong, Y., Li, J., Ding, L., He, X., Liang, K., He, L. and Li, Z. 2017. An improved
894 coupled routing and excess storage (crest) distributed hydrological model and its verification in
895 Ganjiang River Basin, China. Water 9(11), 904.
896 Khakbaz, B., Imam, B., Hsu, K. and Sorooshian, S. 2012. From lumped to distributed via semi-
897 distributed: Calibration strategies for semi-distributed hydrologic models. Journal of Hydrology
898 418, 61-77.
899 Kim, T., Shin, J.-Y., Kim, H. and Heo, J.-H. 2020. Ensemble-Based Neural Network Modeling for
900 Hydrologic Forecasts: Addressing Uncertainty in the Model Structure and Input Variable
901 Selection. Water Resources Research 56(6), e2019WR026262.
902 Kim, T., Shin, J.-Y., Kim, H., Kim, S. and Heo, J.-H. 2019. The use of large-scale climate indices in monthly
903 reservoir inflow forecasting and its application on time series and artificial intelligence models.
904 Water 11(2), 374.
905 Kling, H., Fuchs, M. and Paulin, M. 2012. Runoff conditions in the upper Danube basin under an
906 ensemble of climate change scenarios. Journal of Hydrology 424-425, 264-277.
907 Knoben, W.J., Freer, J.E. and Woods, R.A. 2019. Inherent benchmark or not? Comparing Nash–Sutcliffe
908 and Kling–Gupta efficiency scores. Hydrology and Earth System Sciences 23(10), 4323-4331.
909 Koren, V., Reed, S., Smith, M., Zhang, Z. and Seo, D.-J. 2004. Hydrology laboratory research modeling
910 system (HL-RMS) of the US national weather service. Journal of Hydrology 291(3-4), 297-318.
911 Koutsoyiannis, D. 2003. Climate change, the Hurst phenomenon, and hydrological statistics.
912 Hydrological Sciences Journal 48(1), 3-24.
913 Kratzert, F., Klotz, D., Brenner, C., Schulz, K. and Herrnegger, M. 2018. Rainfall–runoff modelling using
914 long short-term memory (LSTM) networks. Hydrol. Earth Syst. Sci 22(11), 6005-6022.
915 Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A.K., Hochreiter, S. and Nearing, G.S. 2019. Toward
916 improved predictions in ungauged basins: Exploiting the power of machine learning. Water
917 Resources Research 55(12), 11344-11354.
918 Kumar, D., Singh, A., Samui, P. and Jha, R.K. 2019. Forecasting monthly precipitation using sequential
919 modelling. Hydrological Sciences Journal 64(6), 690-700.
920 Laloy, E. and Vrugt, J.A. 2012. High-dimensional posterior exploration of hydrologic models using
921 multiple-try DREAM (ZS) and high-performance computing. Water Resources Research 48(1).
922 Lee, C.J., Rasmussen, P.P., Ziegler, A.C. and Fuller, C.C. 2009 Transport and Sources of Suspended
923 Sediment in the Mill Creek Watershed, Johnson County, Northeast Kansas, 2006-07, U. S.
924 Geological Survey.
925 Leonard, J. and Kramer, M. 1990. Improvement of the backpropagation algorithm for training neural
926 networks. Computers & Chemical Engineering 14(3), 337-341.
927 Li, Z., Yang, Y., Kan, G. and Hong, Y. 2018. Study on the applicability of the Hargreaves potential
928 evapotranspiration estimation method in CREST distributed hydrological model (version 3.0)
929 applications. Water 10(12), 1882.

930 Lin, P., Hopper Jr, L.J., Yang, Z.-L., Lenz, M. and Zeitler, J.W. 2018. Insights into hydrometeorological
931 factors constraining flood prediction skill during the May and October 2015 Texas Hill Country
932 Flood Events. Journal of Hydrometeorology 19(8), 1339-1361.
933 Liu, F., Xu, F. and Yang, S. 2017 A Flood Forecasting Model Based on Deep Learning Algorithm via
934 Integrating Stacked Autoencoders with BP Neural Network, pp. 58-61.
935 Lu, M. and Li, X. 2014. Time scale dependent sensitivities of the XinAnJiang model parameters.
936 Hydrological Research Letters 8(1), 51-56.
937 Ma, Q., Xiong, L., Liu, D., Xu, C.-Y. and Guo, S. 2018. Evaluating the temporal dynamics of uncertainty
938 contribution from satellite precipitation input in rainfall-runoff modeling using the variance
939 decomposition method. Remote Sensing 10(12), 1876.
940 Machiwal, D. and Jha, M.K. (2012) Hydrologic time series analysis: theory and practice, Springer Science
941 & Business Media.
942 Moreda, F., Koren, V., Zhang, Z., Reed, S. and Smith, M. 2006. Parameterization of distributed
943 hydrological models: learning from the experiences of lumped modeling. Journal of Hydrology
944 320(1-2), 218-237.
945 Moriasi, D.N., Arnold, J.G., Van Liew, M.W., Bingner, R.L., Harmel, R.D. and Veith, T.L. 2007. Model
946 evaluation guidelines for systematic quantification of accuracy in watershed simulations.
947 Transactions of the ASABE 50(3), 885-900.
948 Mosavi, A., Ozturk, P. and Chau, K.-w. 2018. Flood Prediction Using Machine Learning Models:
949 Literature Review. Water 10(11), 1536.
950 Naeini, M.R., Analui, B., Gupta, H.V., Duan, Q. and Sorooshian, S. 2019. Three decades of the Shuffled
951 Complex Evolution (SCE-UA) optimization algorithm: Review and applications. Scientia Iranica
952 26(4), 2015-2031.
953 Nash, J.E. and Sutcliffe, J.V. 1970. River flow forecasting through conceptual models part I—A
954 discussion of principles. Journal of hydrology 10(3), 282-290.
955 Nielsen-Gammon, J. 2011 The 2011 Texas drought: a briefing packet for the Texas Legislature.
956 Oubeidillah, A.A., Kao, S.-C., Ashfaq, M., Naz, B.S. and Tootle, G. 2014. A large-scale, high-resolution
957 hydrological model parameter data set for climate change impact assessment for the
958 conterminous US. Hydrology and Earth System Sciences 18(1), 67.
959 Peck, E.L. (1976) Catchment modeling and initial parameter estimation for the National Weather Service
960 river forecast system, Office of Hydrology, National Weather Service.
961 Prat, O. and Nelson, B. 2015. Evaluation of precipitation estimates over CONUS derived from satellite,
962 radar, and rain gauge data sets at daily to annual scales (2002–2012). Hydrology and Earth
963 System Sciences 19(4), 2037.
964 Rahnamay Naeini, M., Yang, T., Tavakoly, A., AghaKouchak, A., Hsu, K. and Sorooshian, S. 2018.
965 Developing a Generalized Model Tree (GMT) framework for simulating reservoir systems.
966 AGUFM 2018, H11U-1747.
967 Rashid, T.A. and Ahmad, H.A. 2016. Lecturer performance system using neural network with Particle
968 Swarm Optimization. Computer Applications in Engineering Education 24(4), 629-638.
969 Rasouli, K., Hsieh, W.W. and Cannon, A.J. 2012. Daily streamflow forecasting by machine learning
970 methods with weather and climate inputs. Journal of Hydrology 414-415, 284-293.
971 Rauf, A.-u. and Ghumman, A.R. 2018. Impact assessment of rainfall-runoff simulations on the flow
972 duration curve of the Upper Indus River—A comparison of data-driven and hydrologic models.
973 Water 10(7), 876.
974 Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J. and Carvalhais, N. 2019. Deep learning
975 and process understanding for data-driven Earth system science. Nature 566(7743), 195-204.
976 Ren-Jun, Z. 1992. The Xinanjiang model applied in China. Journal of Hydrology 135(1), 371-381.

977 Rezaeianzadeh, M., Stein, A., Tabari, H., Abghari, H., Jalalkamali, N., Hosseinipour, E. and Singh, V. 2013.
978 Assessment of a conceptual hydrological model and artificial neural networks for daily outflows
979 forecasting. International journal of environmental science and technology 10(6), 1181-1192.
980 Roodsari, B.K., Chandler, D.G., Kelleher, C. and Kroll, C.N. 2019. A comparison of SAC-SMA and Adaptive
981 Neuro-fuzzy Inference System for real-time flood forecasting in small urban catchments. Journal
982 of Flood Risk Management 12, e12492.
983 Ruggles, R., Brandes, D. and Kney, A.D. 2001 Development of a Geographical Information System for
984 Watershed Research, Information and Education, pp. 1-8.
985 Rumelhart, D.E., Hinton, G.E. and Williams, R.J. 1986. Learning representations by back-propagating
986 errors. nature 323(6088), 533-536.
987 Rumelhart, D.E., Widrow, B. and Lehr, M.A. 1994. The basic ideas in neural networks. Communications
988 of the ACM 37(3), 87-93.
989 Salas, J.D. (1980) Applied modeling of hydrologic time series, Water Resources Publication.
990 Sellars, S. 2018. “Grand Challenges” in Big Data and the Earth Sciences. Bulletin of the American
991 Meteorological Society 99(6), ES95-ES98.
992 Senatore, A., Furnari, L. and Mendicino, G. 2020. Impact of high-resolution sea surface temperature
993 representation on the forecast of small Mediterranean catchments' hydrological responses to
994 heavy precipitation. Hydrology and Earth System Sciences 24(1), 269-291.
995 Shen, C. 2018. A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water
996 Resources Scientists. Water Resources Research 54(11), 8558-8593.
997 Shen, X., Hong, Y., Zhang, K. and Hao, Z. 2017. Refining a distributed linear reservoir routing method to
998 improve performance of the CREST model. Journal of hydrologic engineering 22(3), 04016061.
999 Singh, V.P. 2018. Hydrologic modeling: progress and future directions. Geoscience Letters 5(1), 15.
1000 Singh, V.P. and Woolhiser, D.A. 2002. Mathematical modeling of watershed hydrology. Journal of
1001 hydrologic engineering 7(4), 270-292.
1002 Smith, J.A., Baeck, M.L., Morrison, J.E. and Sturdevant-Rees, P. 2000. Catastrophic rainfall and flooding
1003 in Texas. Journal of Hydrometeorology 1(1), 5-25.
1004 Smith, J.A., Baeck, M.L., Villarini, G. and Krajewski, W.F. 2010. The hydrology and hydrometeorology of
1005 flooding in the Delaware River Basin. Journal of Hydrometeorology 11(4), 841-859.
1006 Smith, M.B., Seo, D.-J., Koren, V.I., Reed, S.M., Zhang, Z., Duan, Q., Moreda, F. and Cong, S. 2004. The
1007 distributed model intercomparison project (DMIP): motivation and experiment design. Journal
1008 of Hydrology 298(1-4), 4-26.
1009 Solomatine, D., See, L. and Abrahart, R.J. (2008), pp. 17-30.
1010 Solomatine, D.P. and Ostfeld, A. 2008. Data-driven modelling: some past experiences and new
1011 approaches. Journal of hydroinformatics 10(1), 3-22.
1012 Sorooshian, S., Duan, Q. and Gupta, V.K. 1993. Calibration of rainfall-runoff models: Application of
1013 global optimization to the Sacramento Soil Moisture Accounting Model. Water resources
1014 research 29(4), 1185-1194.
1015 Sorooshian, S., Hsu, K.-l., Coppola, E., Tomassetti, B., Verdecchia, M. and Visconti, G. (2008) Hydrological
1016 modelling and the water cycle: coupling the atmospheric and hydrological models, Springer
1017 Science & Business Media.
1018 Srivastava, P., McNair, J.N. and Johnson, T.E. 2006. COMPARISON OF PROCESS-BASED AND ARTIFICIAL
1019 NEURAL NETWORK APPROACHES FOR STREAMFLOW MODELING IN AN AGRICULTURAL
1020 WATERSHED 1. JAWRA Journal of the American Water Resources Association 42(3), 545-563.
1021 Srivastava, S. and Lessmann, S. 2018. A comparative study of LSTM neural networks in forecasting day-
1022 ahead global horizontal irradiance with satellite data. Solar Energy 162, 232-247.

1023 Thomsen, B. and Schumann, H.H. 1969 Water resources of the Sycamore Creek watershed, Maricopa
1024 County, Arizona, US Govt. Print. Off.
1025 Tian, J., Liu, J., Wang, Y., Wang, W., Li, C. and Hu, C. 2020. A coupled atmospheric–hydrologic modeling
1026 system with variable grid sizes for rainfall–runoff simulation in semi-humid and semi-arid
1027 watersheds: how does the coupling scale affects the results? Hydrology and Earth System
1028 Sciences 24(8), 3933-3949.
1029 Tokar, A.S. and Markus, M. 2000. Precipitation-runoff modeling using artificial neural networks and
1030 conceptual models. Journal of Hydrologic Engineering 5(2), 156-161.
1031 Valipour, M., Banihabib, M.E. and Behbahani, S.M.R. 2013. Comparison of the ARMA, ARIMA, and the
1032 autoregressive artificial neural network models in forecasting the monthly inflow of Dez dam
1033 reservoir. Journal of hydrology 476, 433-441.
1034 Verdin, J., Funk, C., Senay, G. and Choularton, R. 2005. Climate science and famine early warning.
1035 Philosophical Transactions of the Royal Society B: Biological Sciences 360(1463), 2155-2168.
1036 Vergara, H., Kirstetter, P.-E., Gourley, J.J., Flamig, Z.L., Hong, Y., Arthur, A. and Kolar, R. 2016.
1037 Estimating a-priori kinematic wave model parameters based on regionalization for flash flood
1038 forecasting in the Conterminous United States. Journal of Hydrology 541, 421-433.
1039 Vrugt, J.A., Ter Braak, C.J., Gupta, H.V. and Robinson, B.A. 2009. Equifinality of formal (DREAM) and
1040 informal (GLUE) Bayesian approaches in hydrologic modeling? Stochastic environmental
1041 research and risk assessment 23(7), 1011-1026.
1042 Wang, J., Hong, Y., Li, L., Gourley, J.J., Khan, S.I., Yilmaz, K.K., Adler, R.F., Policelli, F.S., Habib, S. and Irwn,
1043 D. 2011. The coupled routing and excess storage (CREST) distributed hydrological model.
1044 Hydrological sciences journal 56(1), 84-98.
1045 Wang, J., Shi, P., Jiang, P., Hu, J., Qu, S., Chen, X., Chen, Y., Dai, Y. and Xiao, Z. 2017. Application of BP
1046 neural network algorithm in traditional hydrological model for flood forecasting. Water 9(1), 48.
1047 Wang, W.-c., Cheng, C.-T. and Qiu, L. 2009. A Comparison of Performance of Several Artificial
1048 Intelligence Methods for Forecasting Monthly Discharge Time Series. Journal of Hydrology 374,
1049 294-306.
1050 Waseem, M., Mani, N., Andiego, G. and Usman, M. 2017. A review of criteria of fit for hydrological
1051 models. International Research Journal of Engineering and Technology (IRJET) 4(11), 1765-1772.
1052 Widmann, M. and Bretherton, C.S. 2000. Validation of mesoscale precipitation in the NCEP reanalysis
1053 using a new gridcell dataset for the northwestern United States. Journal of Climate 13(11), 1936-
1054 1950.
1055 Wu, W., May, R.J., Maier, H.R. and Dandy, G.C. 2013. A benchmarking approach for comparing data
1056 splitting methods for modeling water resources parameters using artificial neural networks.
1057 Water Resources Research 49(11), 7598-7614.
1058 Xu, D.-m., Wang, W.-c., Chau, K.-w., Cheng, C.-t. and Chen, S.-y. 2013. Comparison of three global
1059 optimization algorithms for calibration of the Xinanjiang model parameters. Journal of
1060 hydroinformatics 15(1), 174-193.
1061 Xu, Y. and Goodacre, R. 2018. On splitting training and validation set: A comparative study of cross-
1062 validation, bootstrap and systematic sampling for estimating the generalization performance of
1063 supervised learning. Journal of Analysis and Testing 2(3), 249-262.
1064 Xue, X., Hong, Y., Limaye, A.S., Gourley, J.J., Huffman, G.J., Khan, S.I., Dorji, C. and Chen, S. 2013.
1065 Statistical and hydrological evaluation of TRMM-based Multi-satellite Precipitation Analysis over
1066 the Wangchu Basin of Bhutan: Are the latest satellite precipitation products 3B42V7 ready for
1067 use in ungauged basins? Journal of Hydrology 499, 91-99.

Yang, T., Asanjan, A.A., Faridzad, M., Hayatbini, N., Gao, X. and Sorooshian, S. 2017a. An enhanced artificial neural network with a shuffled complex evolutionary global optimization with principal component analysis. Information Sciences 418-419, 302-316.

Yang, T., Asanjan, A.A., Welles, E., Gao, X., Sorooshian, S. and Liu, X. 2017b. Developing reservoir monthly inflow forecasts using artificial intelligence and climate phenomenon information. Water Resources Research 53(4), 2786-2812.

Yang, T., Sun, F., Gentine, P., Liu, W., Wang, H., Yin, J., Du, M. and Liu, C. 2019. Evaluation and machine learning improvement of global hydrological model-based flood simulations. Environmental Research Letters 14(11), 114027.

Yao, C., Li, Z.J., Bao, H.J. and Yu, Z.B. 2009. Application of a Developed Grid-Xinanjiang Model to Chinese Watersheds for Flood Forecasting Purpose. Journal of Hydrologic Engineering 14(9), 923-934.

York, J.P., Person, M., Gutowski, W.J. and Winter, T.C. 2002. Putting aquifers into atmospheric simulation models: An example from the Mill Creek Watershed, northeastern Kansas. Advances in Water Resources 25(2), 221-238.

Zealand, C.M., Burn, D.H. and Simonovic, S.P. 1999. Short term streamflow forecasting using artificial neural networks. Journal of hydrology 214(1-4), 32-48.

Zeng, Q., Chen, H., Xu, C.-Y., Jie, M.-X., Chen, J., Guo, S.-L. and Liu, J. 2018. The effect of rain gauge density and distribution on runoff simulation using a lumped hydrological modelling approach. Journal of Hydrology 563, 106-122.

Zhao, R.J. 1992. The Xinanjiang Model Applied in China. J Hydrol 135(1-4), 371-381.

Zhao, R.J., Zhuang, Y. L., Fang, L. R., Liu, X. R., Zhang, Q. S. (ed) (1980) The Xinanjiang model, Hydrological Forecasting Proc., Oxford Symp., IAHS Publication, Wallingford, U.K.

Zhao, W.L., Gentine, P., Reichstein, M., Zhang, Y., Zhou, S., Wen, Y., Lin, C., Li, X. and Qiu, G.Y. 2019. Physics-constrained machine learning of evapotranspiration. Geophysical Research Letters.

Zhijia, L., Penglei, X. and Jiahui, T. 2013. Study of the Xinanjiang model parameter calibration. Journal of Hydrologic Engineering 18(11), 1513-1521.

Zhou, T., Wang, F. and Yang, Z. 2017. Comparative analysis of ANN and SVM models combined with wavelet preprocess for groundwater depth prediction. Water 9(10), 781.

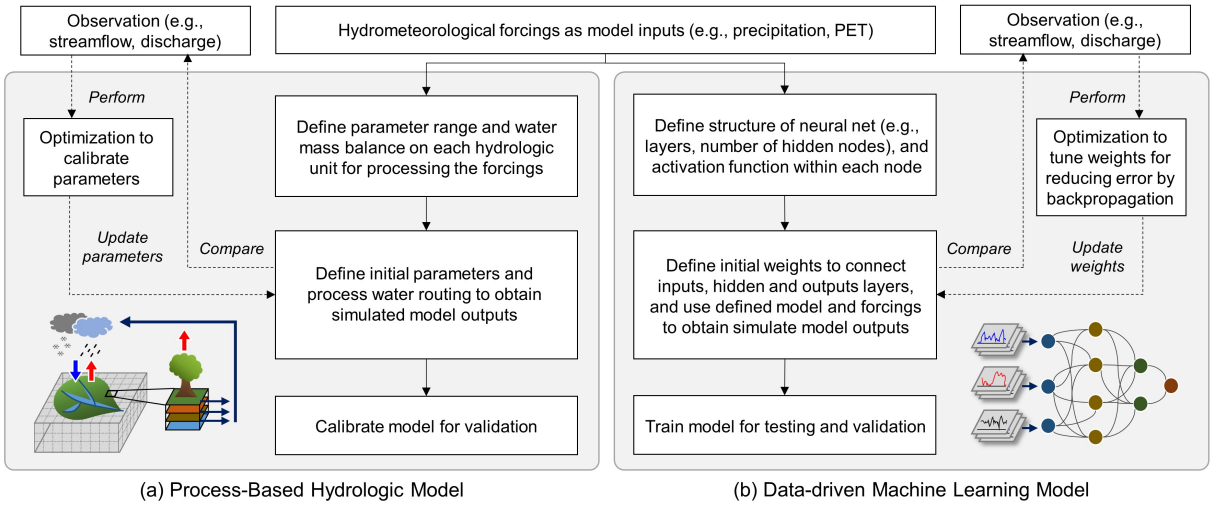


Figure 1 Conceptual comparison between (a) process-based hydrologic models (PHMs) and (b) data-driven machine learning models (DMLs)

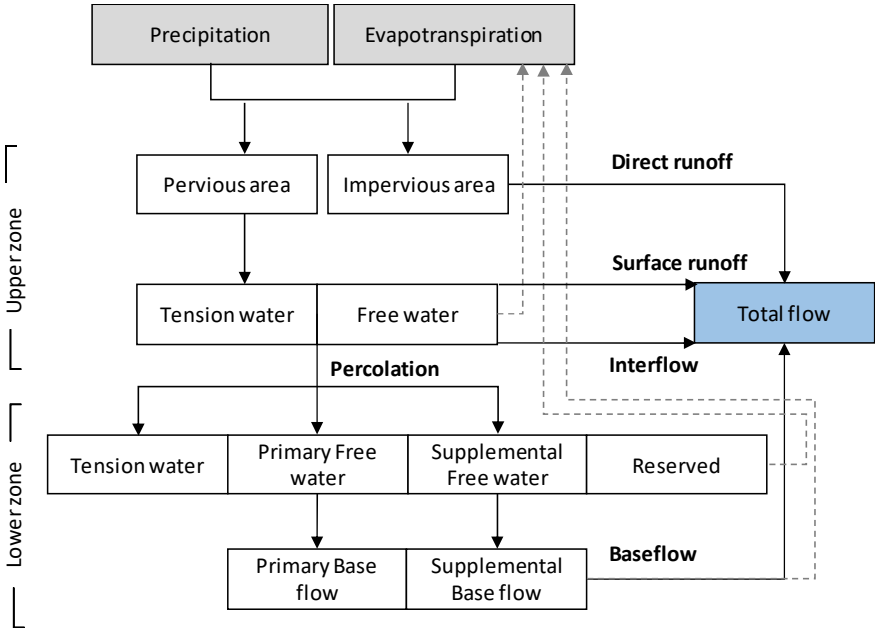


Figure 2 Conceptual diagram of the SAC-SMA model

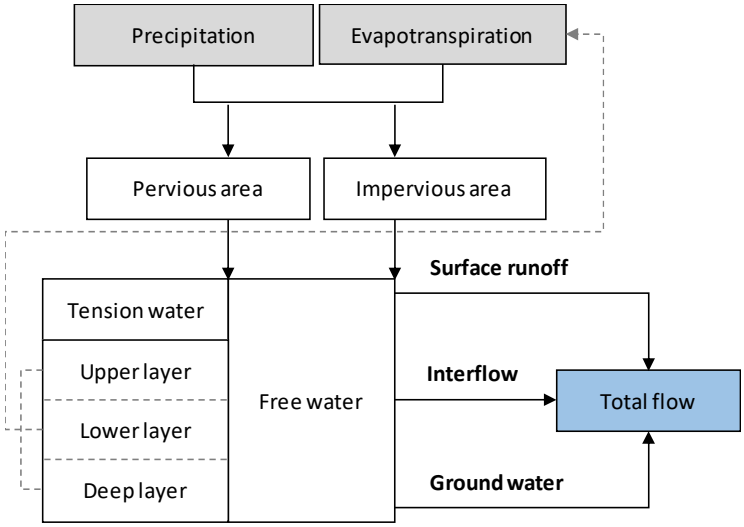


Figure 3 Conceptual diagram of the XAJ model

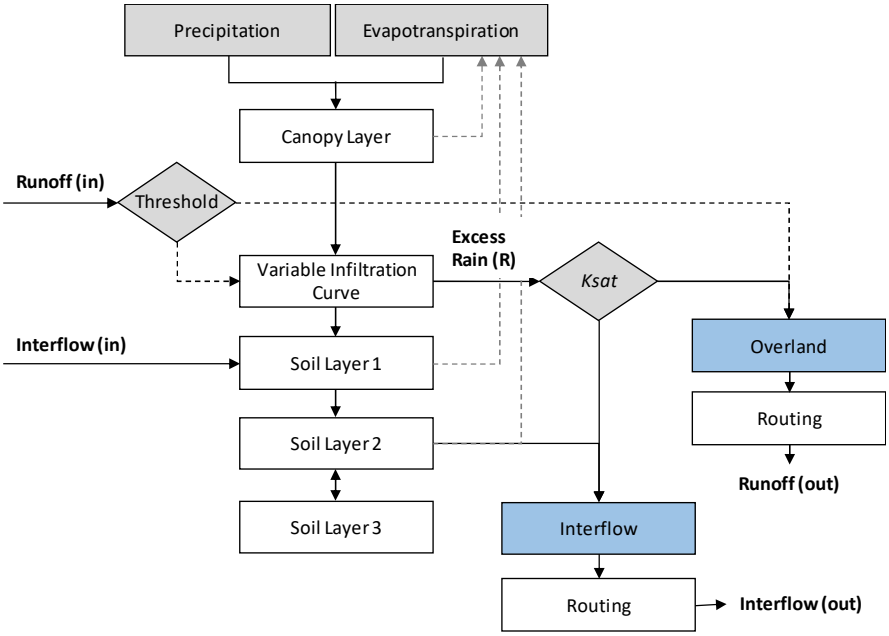
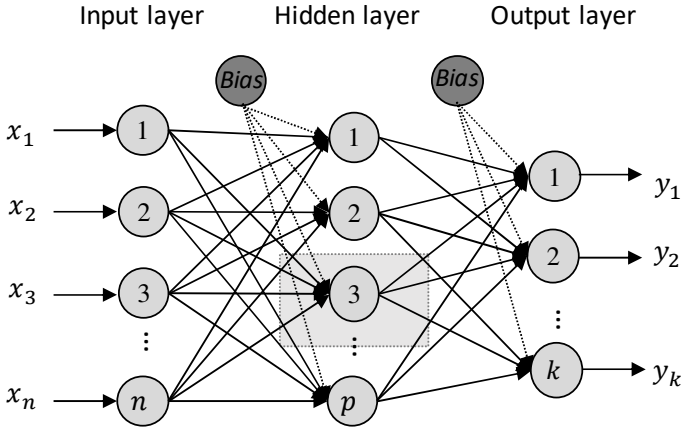


Figure 4 Conceptual diagram of the CREST model

(a) A three-layer feed-forward ANN model structure



(b) The internal operations of the node in the hidden layer

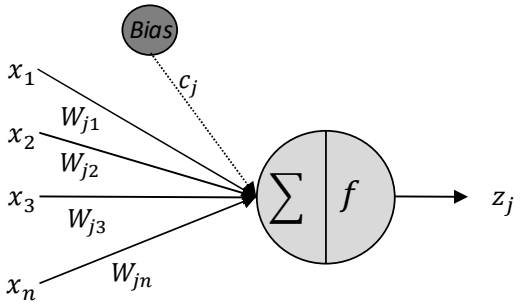
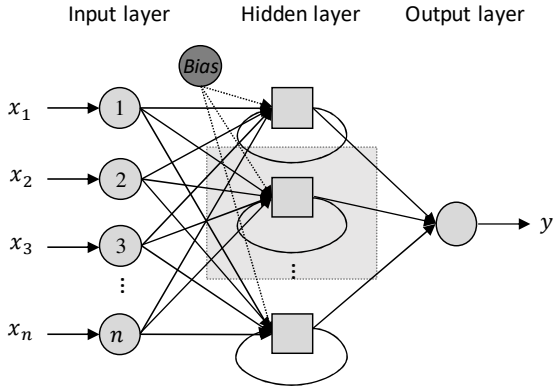


Figure 5 A three-layer feed-forward ANN model structure and the internal operations of the node in the hidden layer

(a) A folded LSTM model structure



(b) The internal operations of the LSTM memory cell in each time step

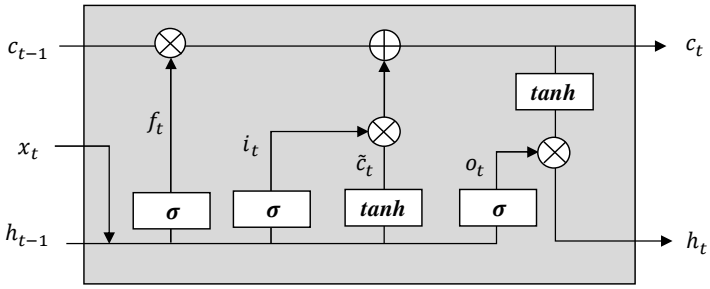


Figure 6 A folded LSTM model structure and the LSTM memory cell in each time step t ($1 \leq t \leq n$)

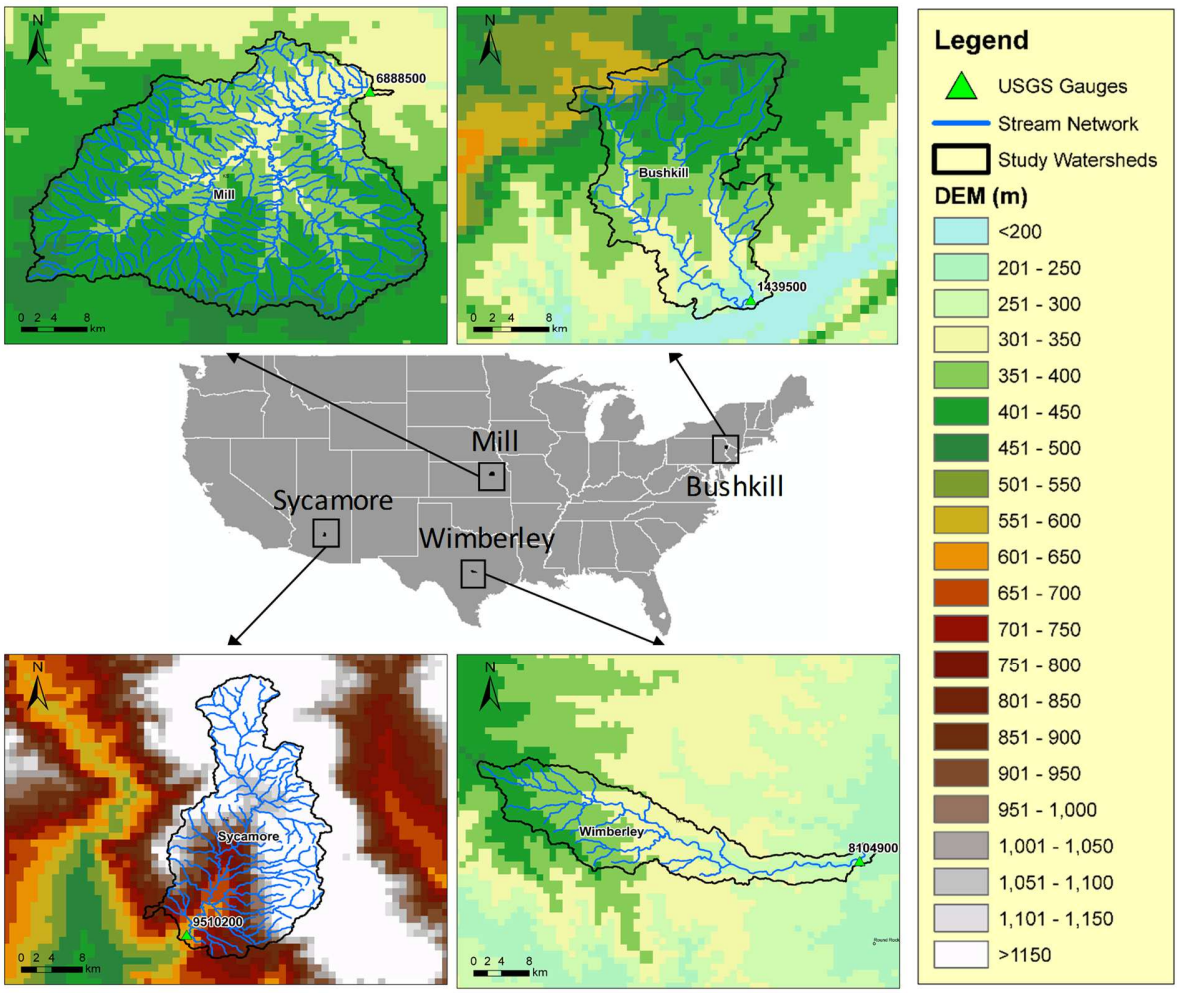


Figure 7 Map of the four watershed over the CONUS (Bushkill is located in Pennsylvania, Mill watershed is located in Kansas, Wimberley watershed is located in Texas, and Sycamore watershed is located in Arizona)

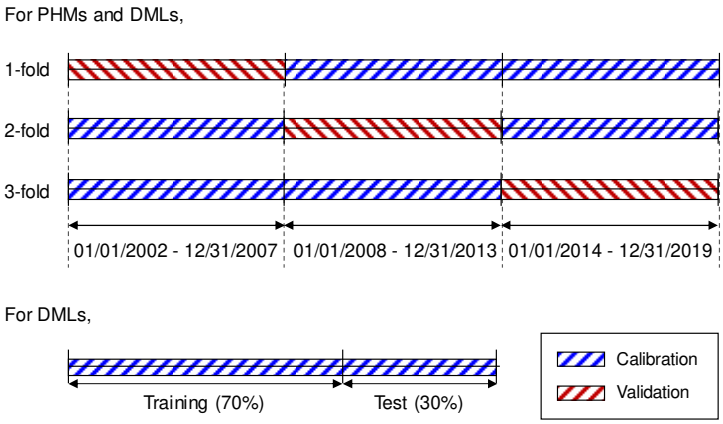


Figure 8 A schematic illustration of k-fold cross-validation (k=3) and data splitting used in this study

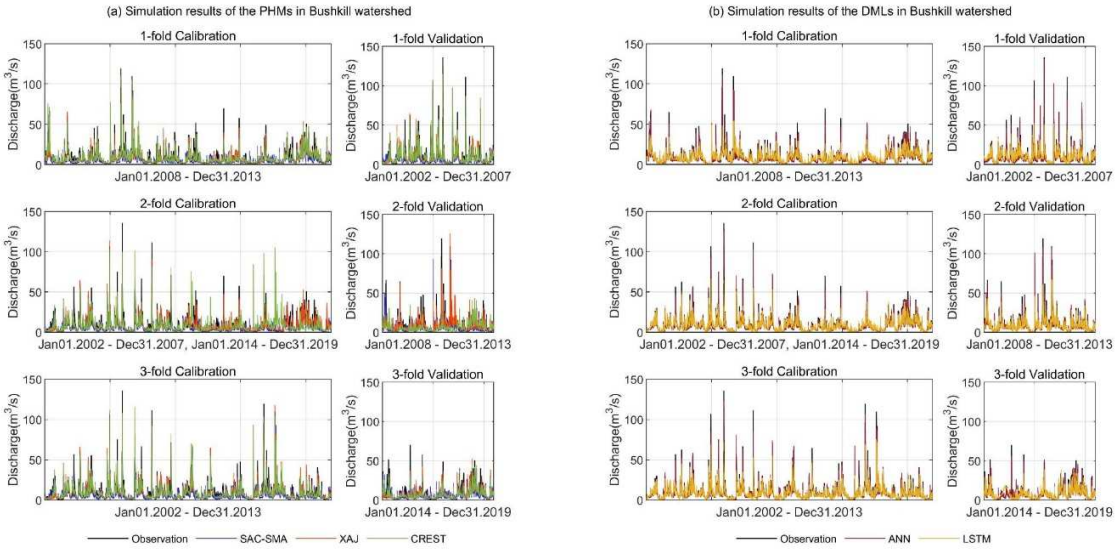


Figure 9 Time series plots of simulation results of the PHMs (left side) and the DMLs (right side) for each k-fold (k=3) in Bushkill watershed (x-axis is the daily time for each calibration and validation set)

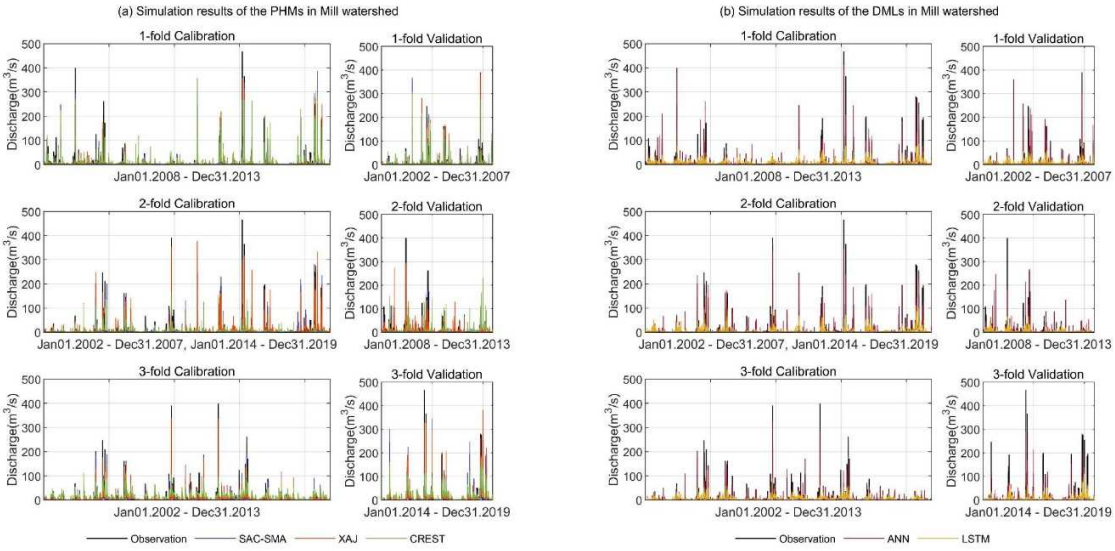


Figure 10 Time series plots of simulation results of the PHMs (left side) and the DMLs (right side) for each k-fold (k=3) in Mill watershed (x-axis is the daily time for each calibration and validation set)

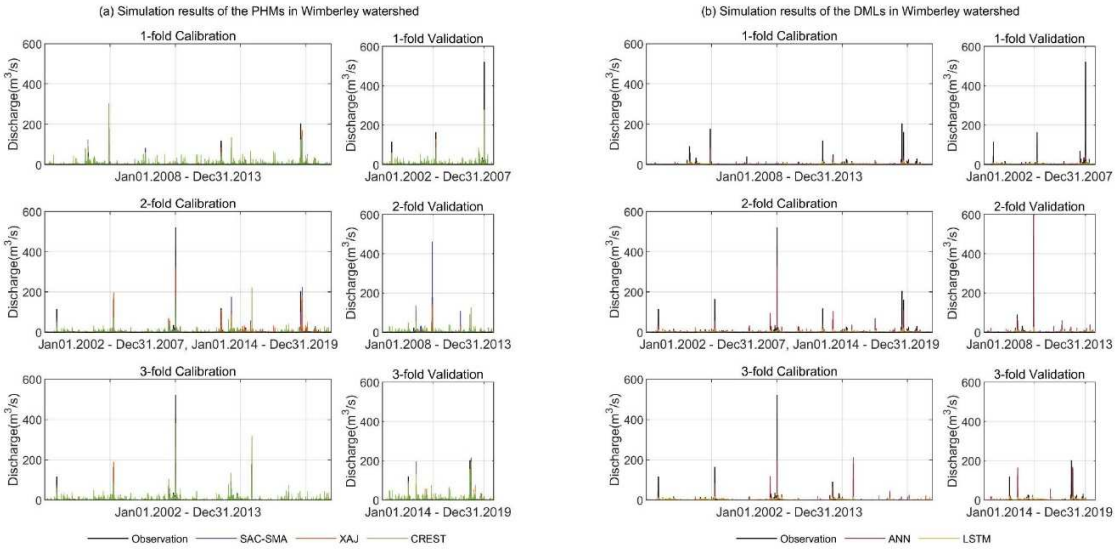


Figure 11 Time series plots of simulation results of the PHMs (left side) and the DMLs (right side) for each k-fold (k=3) in Wimberley watershed (x-axis is the daily time for each calibration and validation set)

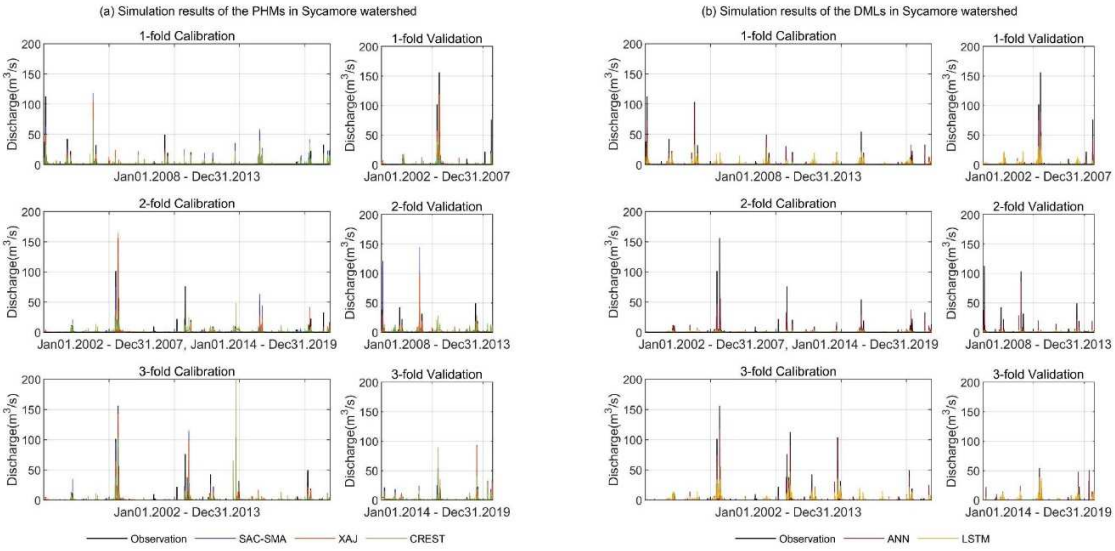


Figure 12 Time series plots of simulation results of the PHMs (left side) and the DMLs (right side) for each k-fold (k=3) in Sycamore watershed (x-axis is the daily time for each calibration and validation set)

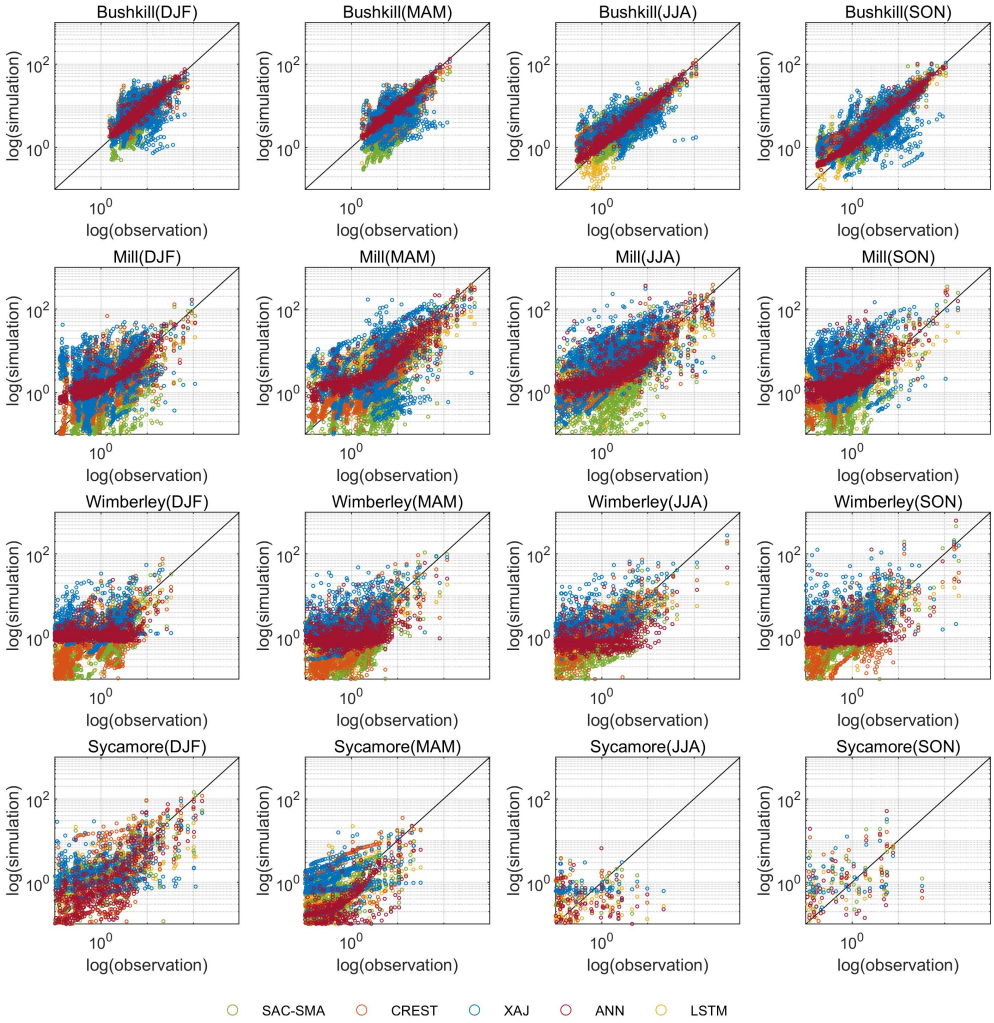


Figure 83 Scatter plots between observed and simulated streamflow during validation periods on a seasonal basis for four watersheds (Bushkill, Mill, Wimberley, and Sycamore from top to bottom; DJF, MAM, JJA, and SON from left to right). Notably, the zeros in both observation and simulation are ignored on the figure

Table 1 SAC-SMA model parameters

No.	Parameter	Description	Units	Range
	Capacity Thresholds (Tunable)			
1	UZTWM	upper zone tension water storage	mm	[10, 300]
2	UZFWM	upper zone free water storage	mm	[5, 150]
3	LZTWM	lower zone tension water storage	mm	[10, 500]
4	LZFPM	lower zone free water primary storage	mm	[10, 1000]
5	LZFSM	lower zone free water supplemental maximum storage	mm	[5, 400]
6	ADIMP	additional impervious area	-	[0, 0.2]
	Recession Parameters (Tunable)			
7	UZK	Upper zone free water lateral depletion rate	day-1	[0.1, 0.75]
8	LZPK	Lower zone primary free water depletion rate	day-1	[0.001, 0.05]
9	LZSK	Lower zone supplemental free water depletion rate	day-1	[0.01, 0.35]
	Percolation and other (Tunable)			
10	ZPERC	Maximum percolation rate	-	[5, 350]
11	REXP	Exponent of the percolation equation	-	[1, 5]
12	PCTIM	Impervious fraction of the watershed area	-	[0, 0.1]
13	PFREE	Fraction percolating from upper to lower zone free water storage	-	[0.0, 0.9]
	Not Optimized (Adjustable)			
14	RIVA	Riparian vegetation area	-	
15	SIDE	Ration of deep recharge to channel base flow	-	
16	RSERV	Fraction lower zone free water not transferable to tension water	-	

Table 2 XAJ model parameters

No.	Parameter	Description	Units	Range
	Evapotranspiration parameters			
1	K	The ratio of potential evapotranspiration to pan evaporation	-	[0.5, 2]
2	UM	Tension water capacity in the upper layer	mm	[0, 20]
3	LM	Tension water capacity in the lower layer	mm	[60, 90]
4	DM	Tension water capacity in the deepest layer	mm	[60, 120]
5	C	The coefficient of deep evapotranspiration		[0, 0.2]
	Runoff production Parameters			
6	B	The exponent of the tension water capacity curve	-	[0.1, 0.4]
7	IM	The ratio of the impervious to the total area of the basin	-	[0.01, 0.1]
	Runoff separation parameters			
8	SM	The areal mean of the free water capacity of surface soil layer	mm	[0, 100]
9	EX	The exponent of the free water capacity curve	-	[1, 1.5]
10	KG	Outflow coefficients of groundwater	-	[0, 0.7]
11	KI	Outflow coefficients of interflow	-	[0, 0.7]
	Runoff concentration parameters			
12	CG	The recession constant of groundwater storage	-	[0.98, 0.998]
13	CS	The recession constant of channel system	-	[0, 1]
14	CI	The recession constant of the lower interflow	-	[0, 0.9]
15	L	Lag time	day	[1, 10]

Table 3 CREST model parameters

No.	Parameter	Description	Units	Range
1	Ksat	The soil saturate hydraulic conductivity	mm/day	[10, 3000]
2	WM	The mean water capacity	mm	[80, 200]
3	B	The exponent of the variable infiltration curve		[0.05, 1.5]
4	IM	Impervious area ratio		[0, 0.2]
5	KE	The factor to convert the potential evapotranspiration to local actual		[0.1, 1,5]
6	TH	Threshold for cumulative drainage area of a channel cell	km^2	[30, 500]
7	UNDER	Interflow speed multiplier		[0.0001, 3]
8	LEAKI	Amount of flow leaking out of interflow reservoir		[0.01, 1]
9	ISU	Initial value of interflow reservoir		[0.0, 0.00001]
10	ALPHA	Multiplier in the $Q = \alpha A^{\beta}$ equation		[0,01, 3]
11	BETA	Exponent in the $Q = \alpha A^{\beta}$ equation		[0,01, 1]
12	ALPHA0	Multiplier in the $Q = \alpha A^{\beta}$ equation for non-channel cells		[0,01, 5]

Table 4 Basic information of the four selected watersheds

Name	Gauge ID (USGS)	Longitude	Latitude	Mean elevation (m)	Basin area (km^2)	Annual mean precipitation (mm)
Bushkill	1439500	-75.0974	41.2099	386	306	1150
Mill	6888500	-96.3137	38.9541	401	843	904
Wimberley	8104900	-97.9712	30.6659	340	348	770
Sycamore	9510200	-111.4729	33.8138	1147	428	295

Table 5 Input scenarios for the PHMs and the DMLs

Input scenario		PHMs			DMLs	
		SAC-SMA	XAJ	CREST	ANN	LSTM
Input scenario	S1	mean precipitation, mean PET			mean precipitation, mean PET	
	S2	-			lagged- mean precipitation and/or mean PET + S1 (Bushkill, Mill: S1 + lagged mean precipitation (1 day) + lagged mean PET(1day); Wimberley, Sycamore: S1+ lagged mean precipitation (1 day))	
	S3	-			lagged- streamflow + S2 (Bushkill, Sycamore: S2 + lagged streamflow (1-, 2-, 3-days); Mill, Wimberley: S2+ lagged streamflow (1 day))	

Table 6 Model performance of simulating the daily streamflow of five employed models using k-fold CV in Bushkill (the statistical measure is the average value of three folds in each calibration and validation set)

Model		RMSE (cms)		CC		NSE		KGE	
		Calibration	Validation	Calibration	Validation	Calibration	Validation	Calibration	Validation
SAC-SMA		4.89	5.06	0.85	0.83	0.70	0.66	0.73	0.73
XAJ		4.18	4.60	0.88	0.85	0.78	0.72	0.82	0.77
CREST		5.62	6.47	0.76	0.60	0.53	0.38	0.74	0.49
ANN	S1	7.45	7.82	0.54	0.47	0.29	0.20	0.34	0.29
	S2	6.81	7.04	0.64	0.60	0.41	0.35	0.46	0.42
	S3	**2.17**	**2.51**	**0.97**	**0.96**	**0.94**	**0.92**	**0.94**	**0.94**
LSTM	S1	4.45	4.88	0.87	0.84	0.75	0.68	0.72	0.69
	S2	4.25	4.48	0.89	0.87	0.77	0.74	0.74	0.72
	S3	3.27	3.43	0.94	0.93	0.86	0.85	0.81	0.80

Table 7 Model performance of simulating the daily streamflow of five employed models using k-fold CV in Mill (the statistical measure is the average value of three folds in each calibration and validation set)

Model		RMSE (cms)		CC		NSE		KGE	
		Calibration	Validation	Calibration	Validation	Calibration	Validation	Calibration	Validation
SAC-SMA		9.30	11.39	0.86	0.77	0.74	0.59	0.72	0.64
XAJ		**8.46**	**9.67**	**0.89**	**0.83**	**0.78**	**0.68**	**0.80**	**0.79**
CREST		15.17	17.84	0.58	0.45	0.21	-0.08	-0.06	-0.08
ANN	S1	12.43	14.10	0.74	0.65	0.54	0.37	0.59	0.50
	S2	12.00	13.52	0.76	0.69	0.57	0.42	0.59	0.52
	S3	9.03	12.11	0.87	0.76	0.75	0.52	0.78	0.67
LSTM	S1	14.08	14.46	0.72	0.66	0.42	0.35	0.33	0.25
	S2	14.19	14.74	0.69	0.61	0.41	0.31	0.35	0.28
	S3	15.19	15.15	0.63	0.56	0.32	0.25	0.38	0.30

Table 8 Model performance of simulating the daily streamflow of five employed models using k-fold CV in Wimberley (the statistical measure is the average value of three folds in each calibration and validation set)

Model		RMSE (cms)		CC		NSE		KGE	
		Calibration	Validation	Calibration	Validation	Calibration	Validation	Calibration	Validation
SAC-SMA		**4.20**	6.77	0.87	**0.82**	**0.74**	0.00	0.66	0.18
XAJ		4.21	**5.41**	**0.87**	0.79	0.74	**0.51**	**0.72**	**0.47**
CREST		6.90	7.21	0.67	0.52	0.26	-0.37	-0.38	-0.99
ANN	S1	5.24	12.43	0.79	0.72	0.58	-5.93	0.55	-0.99
	S2	5.30	12.54	0.78	0.69	0.57	-5.73	0.53	-0.99
	S3	5.49	9.34	0.77	0.67	0.53	-1.06	0.46	-0.28
LSTM	S1	7.70	7.28	0.47	0.49	0.17	0.20	0.07	0.11
	S2	7.84	7.42	0.42	0.43	0.14	0.16	0.04	-0.04
	S3	7.81	7.38	0.42	0.43	0.15	0.18	0.05	0.05

Table 9 Model performance of simulating the daily streamflow of five employed models using k-fold CV in Sycamore (the statistical measure is the average value of three folds in each calibration and validation set)

Model		RMSE (cms)		CC		NSE		KGE	
		Calibration	Validation	Calibration	Validation	Calibration	Validation	Calibration	Validation
SAC-SMA		**2.18**	3.17	**0.84**	**0.72**	**0.71**	0.08	**0.77**	0.25
XAJ		2.36	3.07	0.82	0.70	0.67	0.16	0.70	0.11
CREST		3.33	3.51	0.62	0.52	0.35	0.14	-0.35	-0.63
ANN	S1	2.80	3.45	0.72	0.60	0.50	0.00	0.46	0.10
	S2	3.08	3.13	0.69	0.66	0.44	0.30	0.41	0.33
	S3	2.46	**2.81**	0.81	0.69	0.63	**0.38**	0.61	**0.45**
LSTM	S1	3.28	3.34	0.60	0.59	0.35	0.26	0.21	0.16
	S2	3.13	3.62	0.64	0.40	0.39	0.19	0.21	0.00
	S3	3.32	3.40	0.63	0.56	0.36	0.25	0.30	0.12

Table 10 Mean absolute error of different flow regimes (High-flow, Mid-flow, and Low-flow) in the validation set for each model at four watersheds

Flow regime	Model	Bushkill	Mill	Wimberley	Sycamore
High-flow (≥80%)	SAC-SMA	6.44	10.80	3.89	3.28
	XAJ	6.18	9.11	3.57	4.21
	CREST	9.07	18.87	4.53	3.82
	ANN	**2.35**	**9.03**	4.54	**3.11**
	LSTM	4.05	12.00	**3.36**	3.42
Mid-flow (>20% and <80%)	SAC-SMA	2.70	1.68	**0.42**	**0.27**
	XAJ	2.01	1.96	0.52	0.35
	CREST	2.84	5.11	2.25	1.06
	ANN	**0.61**	**1.13**	0.81	0.29
	LSTM	1.28	2.67	0.87	0.55
Low-flow (≤20%)	SAC-SMA	0.97	**0.65**	**0.14**	**0.10**
	XAJ	1.20	0.77	0.27	0.20
	CREST	1.49	4.36	2.31	0.68
	ANN	**0.34**	1.23	0.91	0.21
	LSTM	0.79	1.93	0.70	0.41

Table 11 Mean and coefficient of variation(CV) of seasonal streamflow at the four watersheds

	Bushkill		Mill		Wimberley		Sycamore	
	Mean (m^3/s)	CV	Mean (m^3/s)	CV	Mean (m^3/s)	CV	Mean (m^3/s)	CV
DJF	9.40	0.82	2.45	2.41	1.11	1.84	1.47	5.51
MAM	10.50	0.96	9.28	3.05	1.63	3.01	0.44	4.08
JJA	5.27	1.49	6.24	3.38	1.30	10.33	0.06	11.14
SON	7.08	1.27	1.56	4.65	1.48	6.56	0.04	9.44