Water Resources Research®

RESEARCH ARTICLE

10.1029/2021WR030993

Key Points:

- Physically based and deep learning models were combined to simulate streamflow in a mountainous karst watershed
- This hybrid modeling approach explicitly accounts for spatially distributed snowmelt
- Convolutional Long Short-Term Memory simulates streamflow with high accuracy and provides insight into spatial hydrologic responses

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

T. Xu,
tianfang.xu@asu.edu

Citation:

Xu, T., Longyang, Q., Tyson, C., Zeng, R., & Neilson, B. T. (2022). Hybrid physically based and deep learning modeling of a snow dominated, mountainous, karst watershed. *Water Resources Research*, *58*, e2021WR030993. https://doi.org/10.1029/2021WR030993

Received 5 AUG 2021
Accepted 22 FEB 2022

Hybrid Physically Based and Deep Learning Modeling of a Snow Dominated, Mountainous, Karst Watershed

Tianfang Xu[1,2], Qianqiu Longyang[1], Conor Tyson[2,3], Ruijie Zeng[1], and Bethany T. Neilson[2]

[1]School of Sustainable Engineering and the Built Environment, Arizona State University, Tempe, AZ, USA, [2]Department of Civil and Environmental Engineering, Utah Water Research Laboratory, Utah State University, Logan, UT, USA, [3]Jordan Valley Water Conservancy District, West Jordan, UT, USA

Abstract Snow dominated mountainous karst watersheds are the primary source of water supply in many areas in the western U.S. and worldwide. These watersheds are typically characterized by complex terrain, spatiotemporally varying snow accumulation and melt processes, and duality of flow and storage dynamics because of the juxtaposition of matrix (micropores and small fissures) and karst conduits. As a result, predicting streamflow from meteorological inputs has been challenging due to the inability of physically based or conceptual hydrologic models to represent these unique characteristics. We present a hybrid modeling approach that integrates a physically based, spatially distributed, snow model with a deep learning karst model. More specifically, the high-resolution snow model captures spatiotemporal variability in snowmelt, and the deep learning model simulates the corresponding response of streamflow as influenced by complex surface and subsurface properties. The deep learning model is based on the Convolutional Long Short-Term Memory (ConvLSTM) architecture capable of handling spatiotemporal recharge patterns and watershed storage dynamics. The hybrid modeling approach is tested on a watershed in northern Utah with seasonal snow cover and variably karstified carbonate bedrock. The hybrid models were able to simulate streamflow at the watershed outlet with high accuracy. The spatial and temporal recharge and discharge patterns learned by the ConvLSTM model were then examined and compared with known hydrogeologic information. Results suggest that ConvLSTM simulates streamflow with higher accuracy than reference models for the study area and provides insight into spatially influenced hydrologic responses that are unavailable within lumped modeling approaches.

1. Introduction

About a quarter of the world's population relies on karst aquifers for drinking water (Hartmann et al., 2014), and one sixth depends on snowmelt water for agriculture and domestic supply (Barnett et al., 2005). In many arid and semi-arid mountainous areas in the western U.S., the northern China, the Middle East, and Mediterranean regions, snow recharged karst aquifers are the primary source of municipal and agricultural water supply (Andreo et al., 2006; de Jong et al., 2008; El-Hakim & Bakalowicz, 2007; Malard et al., 2016; Sweeting, 2012). In mountainous regions in the western U.S., hydrology is dominated by winter snow accumulation and spring melt, with more than 70% of the runoff resulting from snowmelt (Li et al., 2017). Snowpack stores winter precipitation and supplies runoff water throughout the year. While peak runoff typically occurs in late spring or early summer, a large portion of snowmelt water is retained in groundwater stores and sustains summer streamflow. In many regions of the western U.S., summer baseflow, or streamflow contributed by snow recharged aquifers, is the primary source of water for irrigation, where demand peaks in late summer.

How watersheds respond to climate variability has been intensively investigated in catchments throughout the conterminous U.S. (Berghuijs et al., 2014; Harpold et al., 2012; Naz et al., 2016; Sturm et al., 2017), however few studies are focused on karst watersheds (Chen et al., 2018). Hydrologic behavior of karst systems can be characterized by a duality of flow and storage dynamics because of the juxtaposition of matrix (micropores and small fissures) and karst conduits (Hartmann et al., 2014). Rainfall and snowmelt recharge the karst aquifer from a combination of diffuse, slow infiltration into the rock matrix and concentrated, rapid infiltration through sinkholes and vertical fractures (Taylor & Greene, 2008). Subsurface flow travel time varies in orders of magnitude between matrix portion and conduits. Therefore, the recharge and discharge processes in karst watersheds are nonlinear and heterogeneous (Hartmann et al., 2014).

In an effort to understand karst watershed responses and connectivity, various modeling approaches have been applied. Existing spatially distributed karst models discretize the domain in two- or three-dimensional grids, and are capable of representing heterogeneity of hydraulic properties and state variables (Hartmann et al., 2014; Scanlon et al., 2003). However, the use of distributed models has been limited by high computational cost and demand for spatial information about the aquifer property and flow processes. A simpler, more cost-effective alternative is using lumped models that do not explicitly represent spatial variability. A lumped model can be based on one or more conceptualizations of the epikarst, karst conduit, and matrix flow processes (Chang et al., 2017; Mazzilli et al., 2017). The structure of these models (e.g., representation of conduit and matrix storages as linear or nonlinear reservoirs and interaction between reservoirs), is usually determined based on some specific watershed, and model parameters are adjusted through calibration. There are also studies that have constructed data-driven models that inductively infer the transfer function from spatially aggregated input (e.g., precipitation) to output (e.g., discharge) using data fitting techniques such as multivariate regression and machine learning (Z. Li, Wrzesien, et al., 2017). However, these "lumped" machine learning models are not suitable for snow dominated karst aquifers because of the spatially varying snowmelt, rainfall, and recharge processes that influence the hydrologic response and are controlled by complicated meteorological, topographic, and geologic heterogeneity. In addition, existing karst models require site-specific knowledge for configuring the conceptual model structure and parameterization (Chang et al., 2017; Hartmann et al., 2014), which limits the transferability of these models. This creates a clear need for a spatially distributed modeling approach that does not rely on site-specific knowledge about subsurface characterization to set up.

Machine learning techniques are powerful tools for learning complex, nonlinear relations and have been applied in hydrology (Fleming, Garen, et al., 2021; Hsu et al., 1997; Shen et al., 2018; Solomatine & Ostfeld, 2008; Xu & Liang, 2021) and related fields such as agriculture (Liakos et al., 2018), meteorology (McGovern et al., 2017) and remote sensing (Gislason et al., 2006; Lary et al., 2016). For rainfall-runoff modeling in particular, artificial neural networks, decision trees, and kernel methods (e.g., support vector machines, Gaussian process regression) among others achieved comparable or better performance than conceptual hydrologic models in study areas with varied hydrologic regimes (Elshorbagy et al., 2010; Rasouli et al., 2012). Machine learning-based models are not dependent on presumed conceptualization of the hydrologic processes within a given watershed and therefore less prone to model structural error. Successful application of conventional machine learning methods often requires considerable effort and prior knowledge to curate a set of input variables that optimizes the performance of machine learning models, also known as feature engineering (Hastie et al., 2001; LeCun et al., 2015). On the other hand, it has been shown recently that incorporation of domain expertise into feature engineering has the capacity to encourage theory-guided machine learning (Fleming et al., 2015; Fleming & Goodbody, 2019; Karpatne et al., 2017) and improve the interpretability of the trained machine learning model (Fleming, et al., 2021a; 2021b).

Deep learning is a class of machine learning algorithms that typically uses feedforward neural network architectures with more layers than a conventional neural network would have. Deep learning internalizes feature engineering via multiple levels of representation and has been shown to extract information from raw, high-dimension, and large datasets more effectively and objectively than conventional machine learning (LeCun et al., 2015). Two deep network architectures offer great potential for addressing the spatial and temporal complexities of karst aquifers. Long short-term memory (LSTM) networks, a type of recurrent neural network (RNN), has been a popular choice for sequential processes. It is capable of learning long-term dependencies that are typical for hydrologic processes (for example, Fang et al., 2018; Jia et al., 2019). For rainfall-runoff modeling, LSTM achieved accuracy comparable to an established hydrologic model for a wide range of watersheds in the CAMELS data set (Addor et al., 2017; Kratzert et al., 2018, 2019). The second popular type of architecture, namely convolutional networks designed for multi-dimensional data (common in tasks such as image segmentation and object recognition) have also been shown to perform well in hydrologic applications (for example, Anderson & Radic, 2021; Mo et al., 2019; Pan et al., 2019; Sun et al., 2019). As geoscientific applications often deal with spatio-temporal dynamics, there is growing interest in blending LSTM and convolutional networks (Reichstein et al., 2019). The convolutional LSTM (ConvLSTM) architecture combines the strengths of LSTM in representing temporal dynamics and convolutional layers in extracting spatial patterns (Shi et al., 2015). ConvLSTM has been shown effective in capturing spatiotemporal dynamics such as those involved in precipitation nowcasting (Shi et al., 2015).

To further our understanding of karst hydrologic responses, we present a hybrid modeling approach capable of representing the spatial and temporal complexity for a snow dominated mountainous karst watershed in northern

Utah. More specifically, we apply a high-resolution snow model to capture the spatial and temporal variability in snowmelt, and the ConvLSTM model to simulate the response of streamflow to spatially and temporally varying snowmelt and rainfall. The hybrid models can be set up using off-the-shelf data and software tools. Through interpretative analyses, we also demonstrate ConvLSTM learned streamflow responses to spatially varying snowmelt and rainfall.

2. Study Site

The Logan River Watershed is located in the Bear River mountain range of the Rocky Mountains on the Utah-Idaho border. This study focuses on the canyon portion of the Logan River Watershed with an area of 581 km^2, mostly natural land cover (forest, rangeland) with little development. Average annual precipitation and potential evapotranspiration (PET) rates during the study period (1980–2019) are estimated to be 876 and 624 mm, respectively. Most of the runoff volume comes from snow in winter, resulting in a snowmelt dominated hydrograph.

The watershed is underlain by variably karstified carbonate bedrock, with minor siliciclastic intervals, and karst topography. The Ordovician Garden City Formation (limestone) and Silurian Laketown Dolomite host most of karst development in the basin, but all units have the ability to transmit water via dissolution enhanced fractures, faults, bedding planes, and matrix porosity (Spangler, 2001, 2011). An important aquitard in the basin is the Swan Peak Formation (interbedded shales and quartz sandstone) that minimizes vertical groundwater movement between some of the karst layers and intersects the river in multiple places where springs are commonly found. The movement of groundwater is strongly influenced by faults and other structures such as the Logan Peak syncline (Bahr, 2016).

Within the watershed, rainfall and snowmelt recharge comes through (a) sinkholes and pits in meadow beds that are typically developed in high altitude areas, (b) seepage along losing reaches of the basin drainage where the streambed is comprised of permeable fluvioglacial deposits, and (c) diffuse infiltration into ridge slopes (Spangler, 2001). Discharge from the karst aquifer primarily occurs through major (Rick's, Wood Camp, and DeWitt) and minor (Benchmark, Logan Cave) springs along the Logan River. Recent research using mass and flow balances in the Logan River indicates that the bulk of summer low flow is sourced from karst conduits (Neilson et al., 2018). A large portion of Dewitt Spring discharge is diverted to supply drinking water for Logan City before entering the Logan River. Tracer studies have been conducted on the west side of the Logan River to establish subsurface connectivity of karst developed in multiple geologic units and the springs (Spangler, 2001, 2011). These studies indicate flow paths across topographic watershed boundaries occur and in one case, both inter- and intra-basin flow paths join to create Dewitt Springs (Water Canyon and Green Canyon, Figure 1). In order to account for this karst piracy across the topography divide, our study area includes buffering areas outside of the topographically delineated watershed boundary (Figure 1).

Within the vicinity of the Logan River Watershed there are currently 7 SNOTEL (SNOwpack TELemetry) stations in operation (Figure 1). The Franklin Basin and Tony Grove Lake stations have snow water equivalent (SWE) records throughout the study period, while the other five stations have a shorter period of record. Streamflow records of the Logan River since 1953 are provided by USGS gaging station 10109000 (the most downstream station in Figure 1). Upstream of the gaging station, the Highline Canal (USGS 10108400) (Figure 1) diverts a substantial portion of streamflow for agricultural, commercial, and urban irrigation purposes. Monthly use of Dewitt Spring discharge was obtained from Logan City and assumed to occur evenly over each month. The rates of the two diversions were added to streamflow records at USGS 10109000 to calculate the natural streamflow from the canyon and were used to train and validate the deep learning and reference models.

3. Methods

3.1. Physically Based Snow Model

In order to understand the spatiotemporal variability of the snow accumulation and melt processes, we use the Utah Energy Balance (UEB) snow model (Mahat & Tarboton, 2012; Tarboton & Luce, 1996) to calculate the outflow from the base of the snowpack due to melting and/or rainfall (denoted as R hereafter). The UEB model is forced by air temperature, precipitation, wind speed, humidity, and shortwave and longwave radiation. The model simulates the water (SWE, meltwater outflow, and sublimation) and energy (energy content and radi-

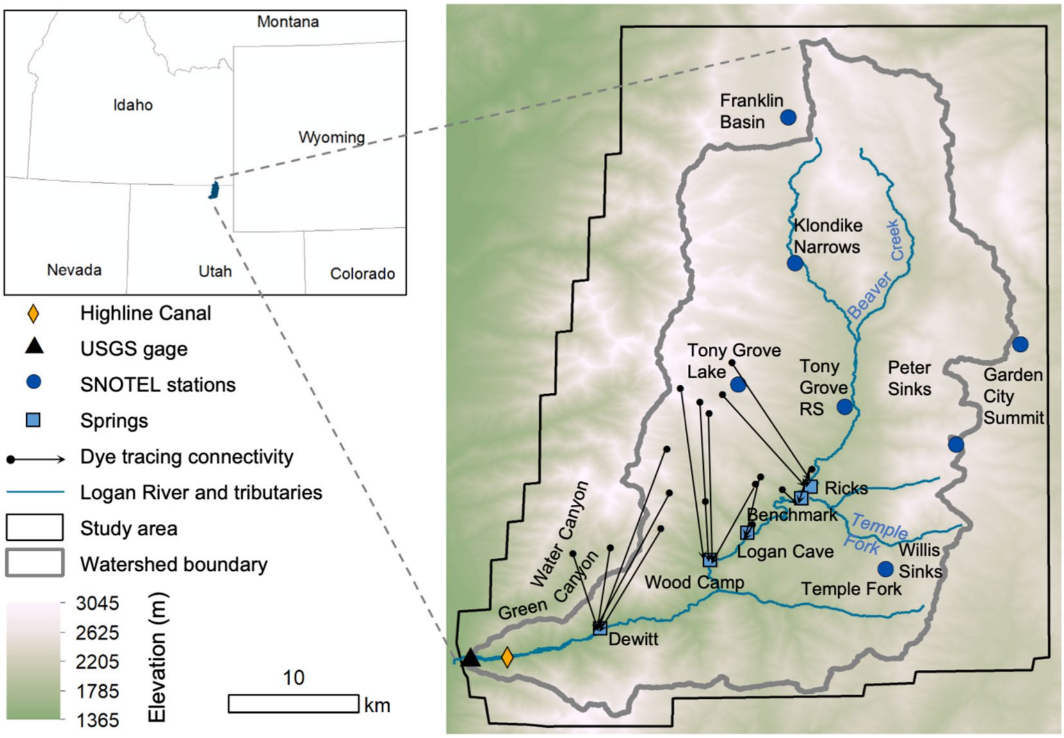

Figure 1. The location (inset) and elevation of the study area. Also shown are the Logan River and its major tributaries (blue), the topographically delineated watershed boundary (gray), the model area boundary (black), SNOTEL stations (blue dots), USGS gage (dark triangle), diversion point of the Highline Canal (diamond), springs (squares). Arrows from dye-injection sites (dark dots) to springs show subsurface connectivity indicated by previous tracer tests (Spangler, 2001, 2011).

ative, sensible, latent, and advective heat exchanges) of the snowpack at time steps sufficient to resolve the diurnal cycle (Tarboton & Luce, 1996). UEB represents the snowpack as a single layer, and therefore is more computationally efficient and has fewer parameters than more complex, multi-layer models such as SNOWPACK (Bartelt & Lehning, 2002) and SNTHERM (Jordan, 1991). Despite being relatively simple, UEB has comparable performance with more complex models (Rutter et al., 2009). In particular, it captures the snowmelt energetics in deep snowpacks (Hood & Hayashi, 2015), which is important for accurately modeling snowmelt (Etchevers et al., 2004).

We ran the UEB model at 100 m resolution and hourly time step for 38 water years (from 1 October 1980 to 30 September 2018). The 100 m resolution allows for characterizing the topographically driven variability of snow accumulation and melt in the mountainous watershed at a reasonable computational cost. Our preliminary results suggest that increasing the grid size to 200 m would cause significant changes in peak SWE to south-facing or north-facing slopes, while reducing the grid size to 50 m caused negligible changes to SWE levels. Leaf area index (LAI) and forest canopy structure (major tree type) parameters were specified using the National Land Cover Database (NLCD) 2011 (Coulston et al., 2012; Yang et al., 2018). Canopy height data was obtained from NASA's LANDFIRE database (Nelson et al., 2013). All of the canopy data were remapped to UEB model grids at 100 resolution. The UEB model was designed to be physically based and transferrable to different locations (Tarboton & Luce, 1996). Therefore, we did not calibrate the UEB model and used recommended parameter values in Mahat and Tarboton (2012), Tarboton and Luce (1996). A list of UEB parameters is provided in Table S1 in Supporting Information S1. The UEB simulated SWE is compared against observations at SNOTEL stations.

In order to generate the forcing data for the UEB model, the North American Land Data Assimilation System (NLDAS-2) Forcing Data set (Xia et al., 2012) was downscaled to 100 m following methods described in Liston and Elder (2006) and Sen Gupta and Tarboton (2016). More specifically, we linearly interpolated NLDAS forcing variables and applied adjustments based on UEB grid elevation, slope, aspect and curvature to account for the effects of complex terrain on spatial distribution of precipitation and other meteorological variables (Tyson, 2021).

Even after orographical adjustment, NLDAS precipitation was found significantly lower than observations at SNOTEL stations (Figure S1 in Supporting Information S1). Therefore, we performed a linear precipitation bias correction following the method used in Sultana et al. (2014) and Sen Gupta and Tarboton (2016). For each SNOTEL station we calculated an annual factor as the ratio between measured annual total precipitation from SNOTEL and orographically adjusted NLDAS at the corresponding UEB grid. The average of the annual factors of all available SNOTEL stations of a given year is applied to all UEB grids to scale the orographically adjusted precipitation. Temperature observations are available at 2–7 SNOTEL stations starting in the year 2000. A comparison with orographically adjusted NLDAS temperature did not reveal significant bias, therefore, bias corrections were not performed for temperature.

3.2. Deep Learning Karst Model

Next, a deep learning model was developed to simulate the hydrologic response of the karst watershed to rainfall and snowmelt (or R) simulated by the UEB model. The deep learning model uses the Convolutional Long Short-Term Memory (ConvLSTM) architecture, which integrates convolution operation into LSTM to capture spatiotemporal dynamics (Shi et al., 2015).

LSTMs are a type of RNN that have proven powerful for learning long-term dependencies in various applications including rainfall-runoff modeling and streamflow forecasting (Kratzert et al., 2018, 2019a, 2019b; Lv et al., 2020; Tennant et al., 2020; Xiang et al., 2020). Each LSTM cell corresponds to one time step, repeats to form N recurrent layers, and retains past information in cell memory. The same weights are shared across time steps. In this study, we implemented the classical LSTM architecture (Hochreiter & Schmidhuber, 1997):

$$\begin{gathered}
i_t = \sigma(W_{xi}\mathbf{x}_t + W_{hi}h_{t-1} + b_i) \\
f_t = \sigma(W_{xf}\mathbf{x}_t + W_{hf}h_{t-1} + b_f) \\
g_t = tanh(W_{xg}\mathbf{x}_t + W_{hg}h_{t-1} + b_g) \\
o_t = \sigma(W_{xo}\mathbf{x}_t + W_{ho}h_{t-1} + b_o) \\
c_t = f_t \odot c_{t-1} + i_t \odot g_t \\
h_t = o_t \odot tanh\,(c_t)
\end{gathered}\tag{1}$$

In the above equations, σ is the logistic sigmoid function, $\odot$ denotes element-wise multiplication, $\mathbf{x}_t$ is the input vector at current time step t, c_t is cell memory, h_t is hidden state, i_t, f_t, o_t are the input, forget, and output gates, respectively, and g_t is the cell input activation vector. W refers to weight matrices, and b denotes bias terms. While i, f, o, h, c, g, b are usually all vectors, here we follow the notation commonly used in deep learning literature. At each time step, the new input is combined with hidden state and cell memory from the previous time step to determine whether the new input will be accumulated to cell memory and whether the past cell memory will be forgotten. The output gate then determines whether the hidden state will be updated with the cell memory.

The classical LSTM applies fully connected input-to-state transition. This is suitable for spatially lumped rainfall-runoff modeling because the input dimension is low (e.g., equals to the number of climate forcings). With spatially distributed inputs, which would have a much higher dimension depending on the spatial discretization, the classical LSTM will have a large number of weights, a lot of which may be redundant. To reduce network complexity and account for spatial correlation structures, ConvLSTM uses convolution operation in input-to-state and state-to-state transitions:

$$\begin{gathered}
i_t = \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + b_i) \\
f_t = \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + b_f) \\
g_t = tanh(W_{xg} * X_t + W_{hg} * H_{t-1} + b_g) \\
o_t = \sigma(W_{xo} * X_t + W_{ho} * X_{t-1} + b_o) \\
C_t = f_t \odot C_{t-1} + i_t \odot g_t \\
H_t = o_t \odot tanh\,(C_t)
\end{gathered}\tag{2}$$

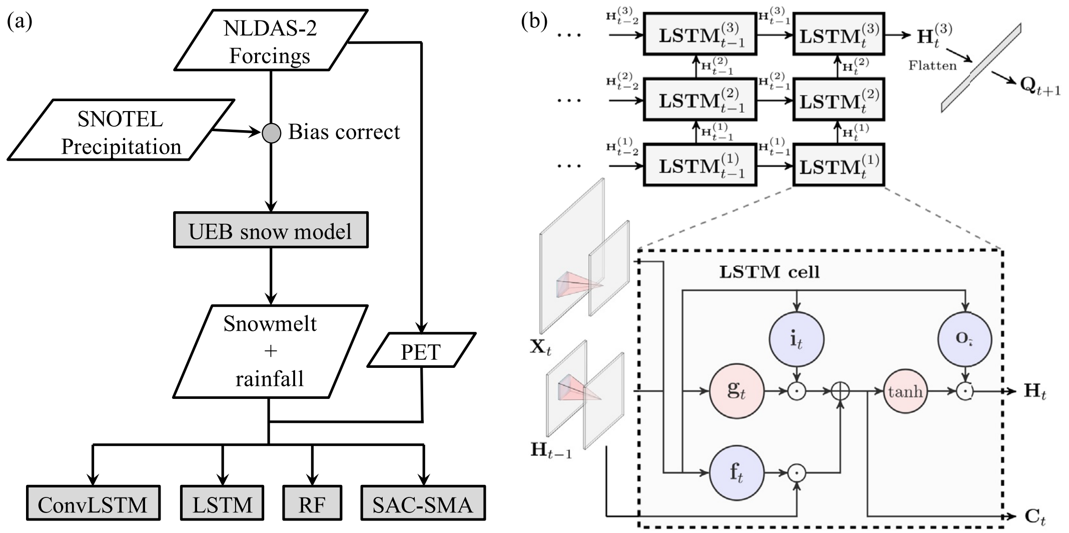

Figure 2. (a) Flowchart of the hybrid modeling approach. Snowmelt and rainfall simulated by the UEB model and PET are used as inputs to the ConvLSTM and three reference models. (b) The architecture of the deep learning karst model with three ConvLSTM layers. At each time step, the input passes through a convolution layer followed by pooling, and the hidden state of the previous step, H_{t-1}, passes through a convolution layer before being used in the gates to calculate current cell memory, C_t(Equation 2). The hidden state of layer 1 is used as inputs by layer 2, and the hidden state of layer 2 is used as inputs by layer 3. Finally, the hidden states of the 3rd layer of ConvLSTM at the current time step, $H_t^{(3)}$, is flattened and fed into a fully connected layer to generate streamflow Q_{t+1}.

In the above equations, X_t is the current input; the input, cell memory (C_t), and hidden state (H_t) are all multi-dimensional arrays and therefore upper cases were used. 2D convolutional operation is denoted by * and defined as below (Goodfellow et al., 2016):

$$S(i, j) = (K * I)(i, j) = \sum_m \sum_n I(i-m, j-n)\, K(m, n), \tag{3}$$

where I is a two-dimensional input, K is the kernel, and S denotes the output or feature map with $S(i, j)$ denoting its ij-th element.

In this study, we stack three ConvLSTM layers (Figure 2) for higher representation power. The input of each time step (1 day in this study) is an image of daily R(snowmelt and rainfall). A second version of the ConvLSTM model is implemented using both R and potential evapotranspiration (PET) as inputs. In both models, layer 1 applies the convolution operation to the input(s) and 2×2 average pooling, that is, aggregation by calculating average of 2×2 pixel values. It also applies convolution to the hidden state from the previous time step. A second ConvLSTM layer then takes the hidden state $H_t^{(1)}$ (superscript denotes layer 1) as inputs and calculates the hidden state of layer 2, $H_t^{(2)}$, which is used as inputs by layer 3. Finally, a fully connected layer processes the hidden states from layer 3 to generate streamflow of the following day, Q_{t+1}. All convolution operation uses 3×3 kernels with 10 output channels and padding. In total, the one-input and two-input ConvLSTM models have 19,691 and 20,051 learnable parameters, respectively. The ConvLSTM model is implemented in Apache MxNet, an open-source deep learning framework (Chen et al., 2015).

Similar to Kratzert et al. (2018), a lookback of 365 days is used in this study. More precisely, the model uses $\{R_{t-365}, R_{t-364}, ..., R_t\}$ (and $\{PET_{t-365}, PET_{t-364}, ..., PET_t\}$ for the two-input model) to simulate streamflow of the following day, Q_{t+1}. We assume that 365 days should capture most of the streamflow variability at a reasonable computational cost based on the high level of karstification of the Logan River Watershed. Previous tracer studies have suggested maximum groundwater travel times of 8–35 days from losing streams in high elevation to major springs within the watershed (Spangler, 2001, 2011). Longer term dependency due to watershed storage is handled by a two-stage training strategy that uses historical information to specify the initial cell memory and state at day $t - 365$.

Directly using the high resolution (100 m) UEB simulation results as input would lead to high computational and memory cost as well as overfitting when training the ConvLSTM models. Therefore, we aggregated the UEB simulated snowmelt and rainfall on a given day (R_t) to 1.6 km resolution by taking the average of R_t across all UEB grids within a 1.6 km by 1.6 km grid; this led to a matrix with reduced size (27 × 18) daily. Although coarsening is necessary due to the computational constraints, using a 100 m resolution UEB model enables characterizing the nonlinear topography effects on snowmelt intensity and timing, which is important for accurately simulating streamflow response. For example, within a 1.6 km by 1.6 km grid, deeper snowpack at high elevation and north-facing slopes melts late, while shallower snowpack at low elevation and south-facing slopes melts earlier. The prolonged snowmelt period resulting from a non-uniform snowpack is captured by the aggregated snowmelt (Figure S5 in Supporting Information S1). However, UEB simulation using a 1.6 km by 1.6 km grid would lead to a uniform snowpack and a shorter melt period than the non-uniform snowpack simulated at 100 m resolution. Next, PET rates were calculated using the Priestley-Taylor method (Priestley & Taylor, 1972) based on orographically adjusted climate forcings. Similar to the snowmelt, the PET rates were coarsened to a 1.6 km by 1.6 km resolution. All input and output data were linearly scaled to the range of (0,1) for the ConvLSTM model.

3.3. Reference Models

In addition to the ConvLSTM models, we use the UEB simulated snowmelt and rainfall and PET (both coarsened to 1.6 km) to drive three reference models to evaluate the strengths and weaknesses of more common modeling approaches.

3.3.1. LSTM

In order to assess the added value of the spatial information, we compare the performance between ConvLSTM and LSTM models, with the latter used as a lumped model that takes as input the spatially averaged daily R (and PET in the case of a two-input model). Therefore, the input of day t is either a scalar ($x_t = R_t$) or a two-element vector ($\mathbf{x}_t = [R_t,\ PET_t]$). Similar to the ConvLSTM models, a lookback period of 365 days is used. Three LSTM layers are stacked, each layer having 20 hidden units. The one-input and two-input LSTM models have 8,581 and 8,661 learnable parameters, respectively.

3.3.2. Random Forest

Using conventional machine learning methods for rainfall-runoff modeling has been intensively investigated (for example, Rasouli et al., 2012; Solomatine & Ostfeld, 2008). Conventional machine learning techniques require smaller training datasets than deep neural networks. However, these techniques are not as powerful in terms of extracting information from natural data in their raw form (LeCun et al., 2015). Therefore, a considerable amount of effort needs to be taken for feature engineering, or creating a suitable set of features (inputs) from the raw data.

The Random Forest (RF) algorithm has been used successfully in various hydrologic applications (Naghibi et al., 2016; Xu et al., 2017) and other fields such as meteorology (Cloke & Pappenberger, 2008; He et al., 2016). A RF model consists of an ensemble of Classification and Regression Trees (CARTs) and outputs the mean prediction of individual trees. A CART tree is composed of a sequence of binary splits. Each split partitions the input space into two regions, and a constant value is fitted to each region. This recursive process stops when the number of observations at the terminal nodes (leaves) is fewer than a threshold (leaf size). As such, a CART estimates a piecewise constant function of the input variables. The RF algorithm trains each tree using a bootstrap sample (sample with replacement) of the training data set; the data points left out can be used to calculate out-of-bag error as an estimate of generalization error. At each binary split, the RF algorithm samples a random subset of input features and identifies the best splitting feature within the subset by minimizing the sum of squares error. The size of the random subset is conventionally set to a third of the total number of input features (Svetnik et al., 2003), and the leaf size is tuned by minimizing the out-of-bag error. RF also measures the importance of input variables by calculating the increase in out-of-bag error when a given input is perturbed. Because of the randomness introduced into the training process, RF is believed to be more robust than CARTs (Breiman, 2001; Hastie et al., 2001).

The RF structure does not inherently account for temporal dependency in that the output at time step t is completely determined by inputs at t. Theoretically, one can augment the inputs with historical information, that is, $\mathbf{x}_t = \{R_{t-365}, R_{t-364}, \ldots, R_t; PET_{t-365}, PET_{t-364}, \ldots, PET_t\}$. However, the resulting high dimensionality will likely cause inferior generalization performance. Therefore, feature engineering was performed to create new features that contain information from the past, including dryspell (number of days since last day of effective rainfall or snowmelt), lagged moving window average time series of R, and accumulated R and PET (Table S2 in Supporting Information S1). These features were selected based on correlation analyses between streamflow and time series with varying lags.

3.3.3. SAC-SMA

Given the broad-spread application of the Sacramento Soil Moisture Accounting (SAC-SMA) model (Burnash et al., 1973), it was selected as the conceptual hydrologic model. The SAC-SMA model is routinely used for rainfall-runoff forecasting including in the U.S. National Weather Service River Forecast System (Demargne et al., 2014) and has been used as a benchmark for data-driven methods (Kratzert et al., 2018; Newman et al., 2015). Similar to Newman et al. (2015), streamflow routing was performed using a two-parameter unit hydrograph (Nash, 1957; Table S3 in Supporting Information S1). The unit hydrograph shape and scale parameters, the scaling factor of PET estimated by the Priestley-Taylor method, and 16 SAC-SMA parameters (Table S3 in Supporting Information S1) were calibrated. Although not designed for karst watersheds, the SAC-SMA model structure allows representing a variety of recession behaviors found in nature. In particular, it uses a primary lower zone free water reservoir that drains slowly to simulate long-term sustaining baseflow, and a supplementary lower zone free water reservoir that drains faster to simulate baseflow after rainfall/snowmelt events. The parallel reservoirs resemble the structure of some conceptual karst models (Butscher & Huggenberger, 2008; Fleury et al., 2007) and may be able to represent the duality of storage and flow in karst conduits and matrix.

3.4. Model Training, Test, and Interpretative Analyses

We use streamflow data and UEB simulation results for 38 water years (WY), October 1980 to September 2018, to train/calibrate and validate the ConvLSTM and three reference models. For the ConvLSTM and LSTM models, the study period was divided into three segments: data during WY 1981–2007 are randomly partitioned into training and validation datasets (ratio = 8:1). During training, the generalization errors of the ConvLSTM and LSTM models are monitored on the independent validation data set, and training stops when the generalization error begins to rise. The training process seeks to minimize the sum of mean-square-error (MSE) and an L_2 norm of the learnable parameters as a penalty of model complexity. The weight of penalty is determined based on the validation loss. In addition, dropout is performed at a rate of 30% (Hinton et al., 2012). Training was performed once with cell memory and hidden state initialized to zero. We then run the trained model from the beginning of the study period to calculate the cell memory and state of each day, which carry historical information beyond the 1-year lookback. Next, the ConvLSTM and LSTM models are initialized with the calculated cell memory and state, and the training resumes until early stopping. The random forest model is trained using data of WY 1981–2007 by minimizing MSE. The SAC-SMA model uses WY 1981–1984 as a spin-up period and WY 1985–2007 for calibration. Calibration is performed by minimizing MSE using DREAM-ZS (Vrugt et al., 2009). We use streamflow during the test period (WY 2008–2018) to evaluate the performance of the trained/calibrated models based on four measures, including percent bias (PBIAS, Gupta et al., 1999), root-mean-square error (RMSE), Nash-Sutcliff efficiency (NSE), and Kling-Gupta efficiency (KGE) (Gupta et al., 2009).

Lack of physical feasibility has long been recognized as a primary drawback of machine learning and deep learning when being applied to hydrologic problems. Therefore, we examined how the trained/calibrated models represent the hydrologic behavior of the study watershed. We first analyzed the streamflow response to snowmelt pulses as simulated by all the models. A 4-year spin up period was first run with inputs equal to long-term average. Snowmelt pulses are then introduced by applying spatially uniform snowmelt at two different rates, 20 mm (lasting for 1 day) and 5 mm (lasting for 4 days). The resulting streamflow was then deducted by streamflow simulated with no pulse to calculate the difference (ΔQ) as a function of time.

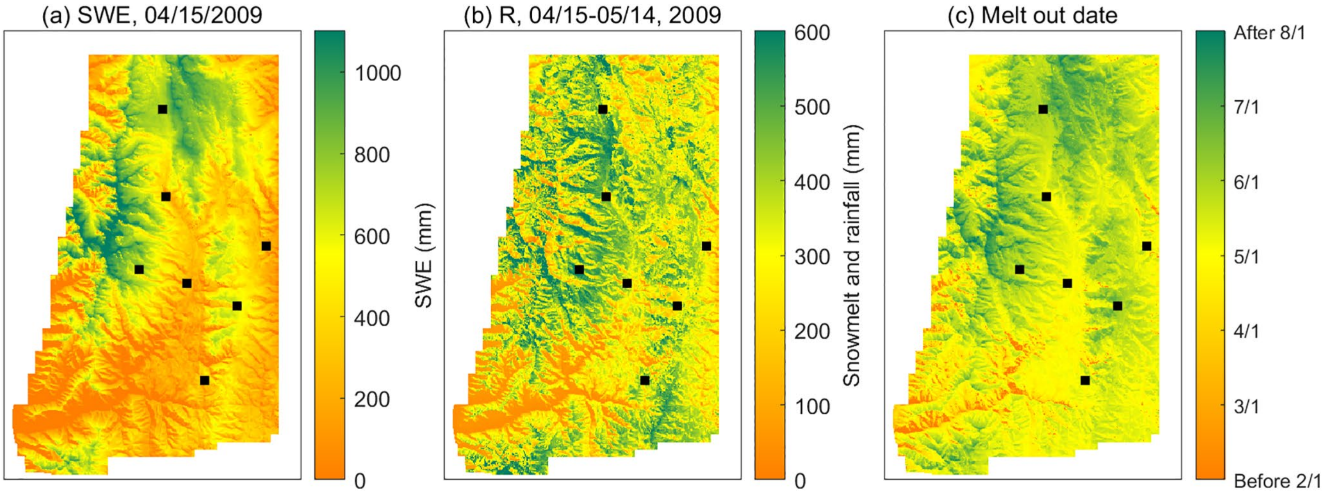

Figure 3. (a) Snow water equivalent on 15 Apr. 2009; (b) Accumulated snowmelt and rainfall (R) during 04/15/2009–05/14/2009; (c) Date when the snowpack completely melted in 2009. Black squares are the locations of SNOTEL stations (Figure 1).

To understand whether the ConvLSTM model provides a spatial understanding of recharge-discharge processes that the other models do not capture, we also performed a sensitivity analysis to examine the streamflow response to spatial snowmelt and rainfall (R) learned by ConvLSTM. Because of the prevalence of fast, concentrated recharge through sinkholes and fractures in the Logan River Watershed, we use R as the surrogate for recharge. More specifically, for each of the 27 × 18 grids we calculate the change in simulated streamflow on day T ($\Delta Q_{T,i,j,m,n}$) induced by adding a perturbation to snowmelt and rainfall of the i, j-th grid on days $T - m, \ldots, T - n$. As such, a high sensitivity value is expected for recharge areas. Because the sensitivity is dependent on the status of the watershed as well as past climate conditions, we calculate the spatial sensitivity for streamflow on two separate days during spring runoff and summer recession, respectively.

4. Results

4.1. UEB Simulated SWE

We use the UEB model to simulate hourly SWE at a 100 m resolution across the study domain (Figure 3). The simulated SWE is compared with observations at SNOTEL stations (Figure 4). Overall, the UEB simulation underestimates SWE at the SNOTEL stations for most of the simulated years. The underestimation at the Tony Grove Lake and USU Doc Daniels stations can largely be explained by a discrepancy in precipitation. At the two stations the orographically adjusted NLDAS precipitation is substantially lower than SNOTEL measurements (Figure S1 in Supporting Information S1), and the spatially averaged annual bias correction factor did not fully remove the bias. In addition, the SNOTEL precipitation measurements used for bias correction are subject to wind-induced undercatch (Avanzi et al., 2014; Groisman & Easterling, 1994), while SWE may be overestimated due to snow drifting (Meyer et al., 2012) and interception redistribution. Because SNOTEL stations are often located in small forest clearings, snow intercepted by surrounding canopy may be redistributed to the snow pillow (Mahat & Tarboton, 2014). Therefore, the point-scale SWE measurements at the SNOTEL sites may not be representative of the corresponding UEB grid and tend to be higher. In addition to the uncertainties due to spatial scale mismatch (Chen et al., 2014), point-based SWE measurements may be subject to errors up to 30% (Schlögl et al., 2016). Given the presence of precipitation and SWE measurement biases, the UEB simulation captures the snow accumulation and melt processes reasonably well. Calibration tests suggested that tuning the UEB parameters within the physically feasible range did not substantially alter the UEB simulation results toward observations at SNOTEL stations.

4.2. Streamflow Simulated by ConvLSTM and Reference Models

As described in Section 3, two versions of ConvLSTM and LSTM models were implemented that use R both with and without PET. Including PET as a second input led to negligible differences in the performance metrics

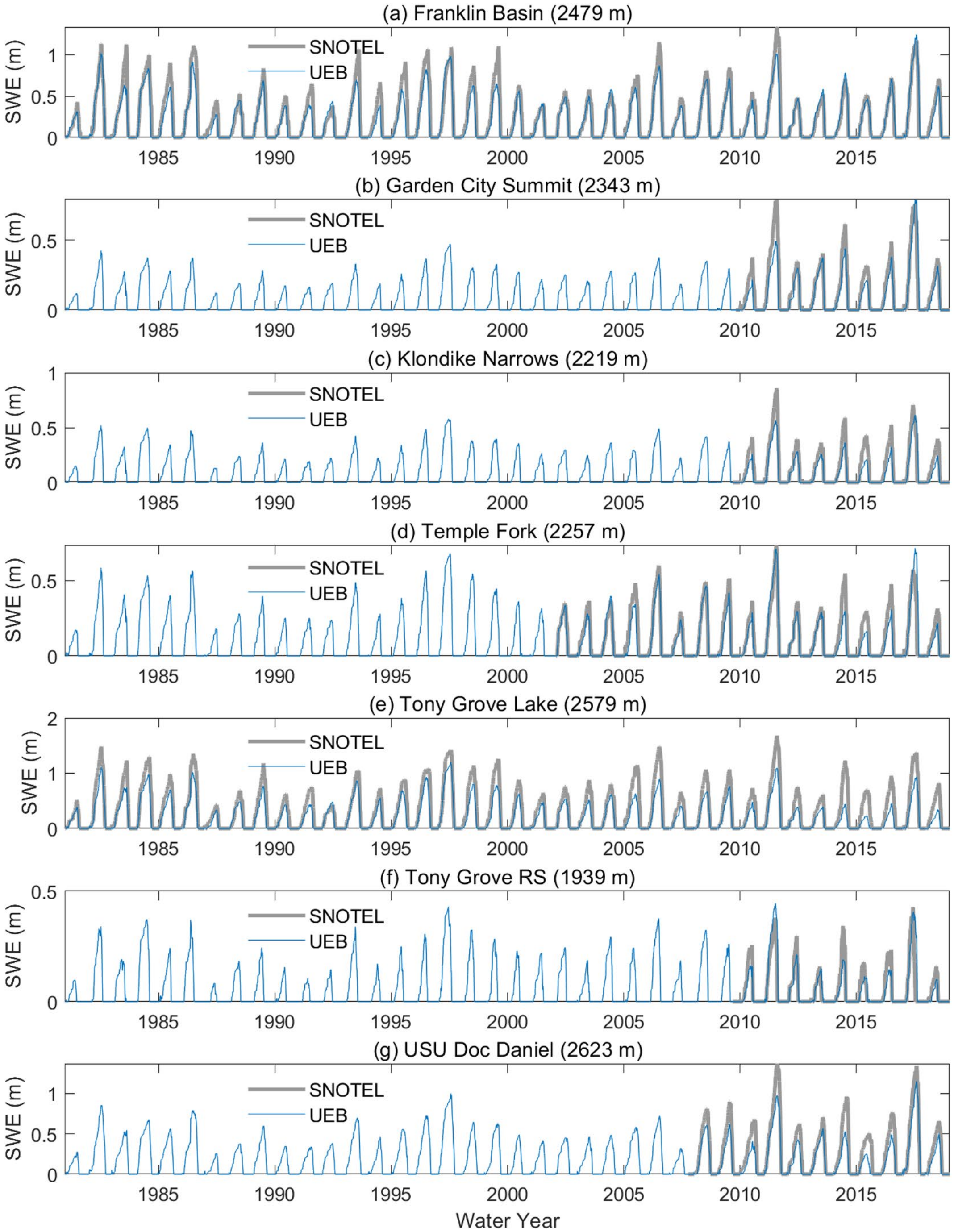

Figure 4. Daily snow Water Equivalent (SWE) observations at seven SNOTEL stations (Figure 1), with numbers in parentheses the elevation of each station. Also shown is the UEB simulated SWE (gray) of the model grid in which a station is located at the end of each day.

but spurious patterns being learned by the ConvLSTM and LSTM models. This confounding issue is further discussed in Section 5. Therefore, only the results obtained from one-input ConvLSTM and LSTM models forced by snowmelt and rainfall are presented. The RF and SAC-SMA models use both *R* and PET.

All models were able to simulate streamflow reasonably well (Table 1, Figures 5 and 6). Given that a PBIAS approaching zero, lower RMSE, and NSE and KGE approaching one indicate a better match between simulations and observations, the ConvLSTM model performed the best during the test period (WY 2008–2018). The

Table 1
Percent Bias (PBIAS), Root-Mean-Square Error (RMSE), Nash Sutcliff Efficiency (NSE), and Kling-Gupta Efficiency (KGE) of the Modified ConvLSTM, LSTM, Random Forest, and SAC-SMA Models During the training and Test Periods

	Train/Calibrate (1981–2007)				Test (2008–2018)			
Model	PBIAS (%)	RMSE (mm/day)	NSE	KGE	PBIAS (%)	RMSE (mm/day)	NSE	KGE
ConvLSTM	1.0	0.39	0.86	0.92	3.2	0.38	0.87	0.93
LSTM	−1.0	0.42	0.84	0.87	−4.5	0.48	0.79	0.89
Random Forest	−0.26	0.31	0.93	0.91	−13	0.49	0.78	0.81
SAC-SMA	0.21	0.38	0.87	0.91	−8.9	0.46	0.80	0.87

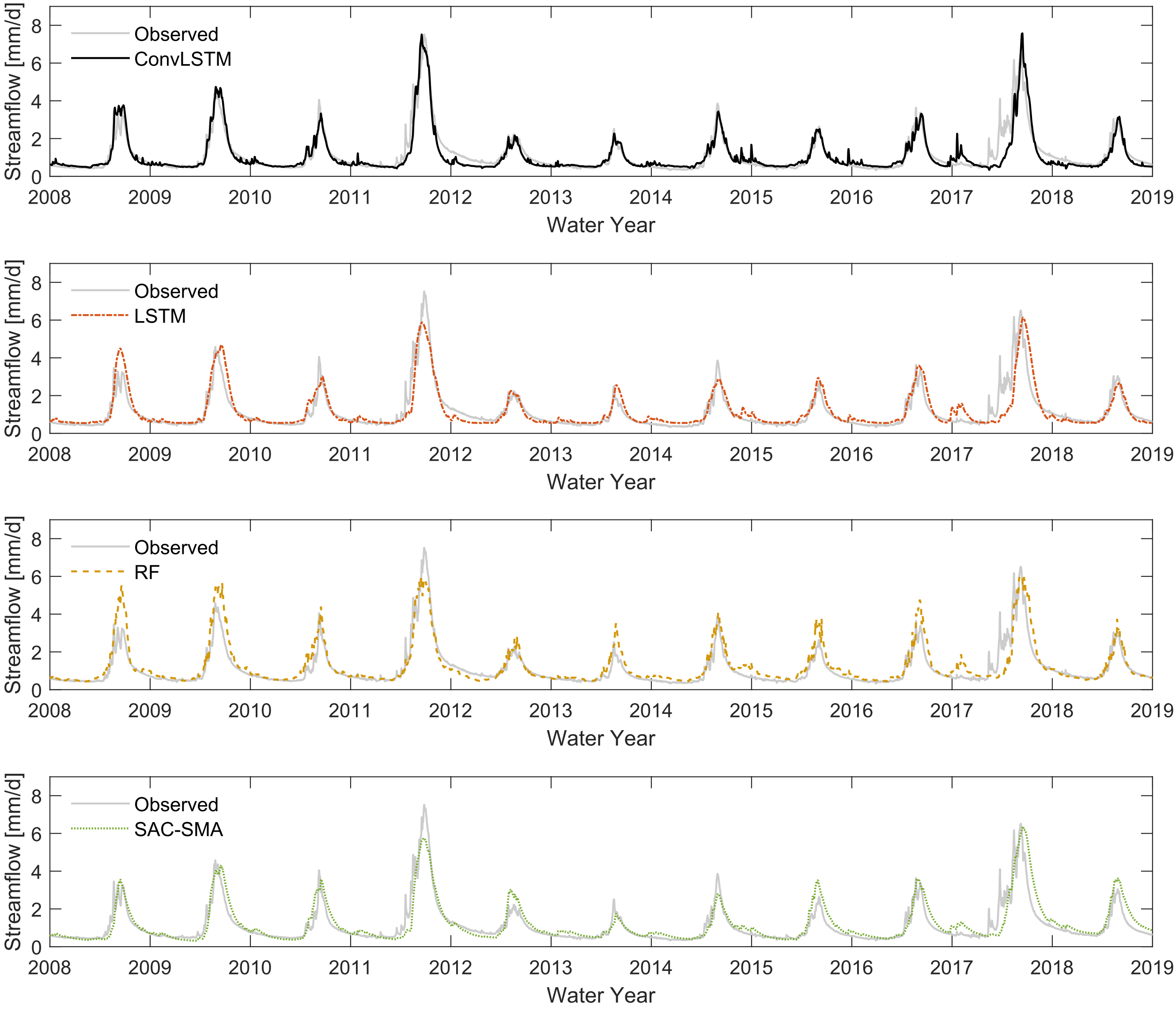

Figure 5. Observed streamflow and simulations (thick gray lines) given by ConvLSTM (a), LSTM (b), RF (c), and SAC-SMA (d) during the testing period (WY 2008–2018). Streamflow is normalized by the topographically delineated watershed area.

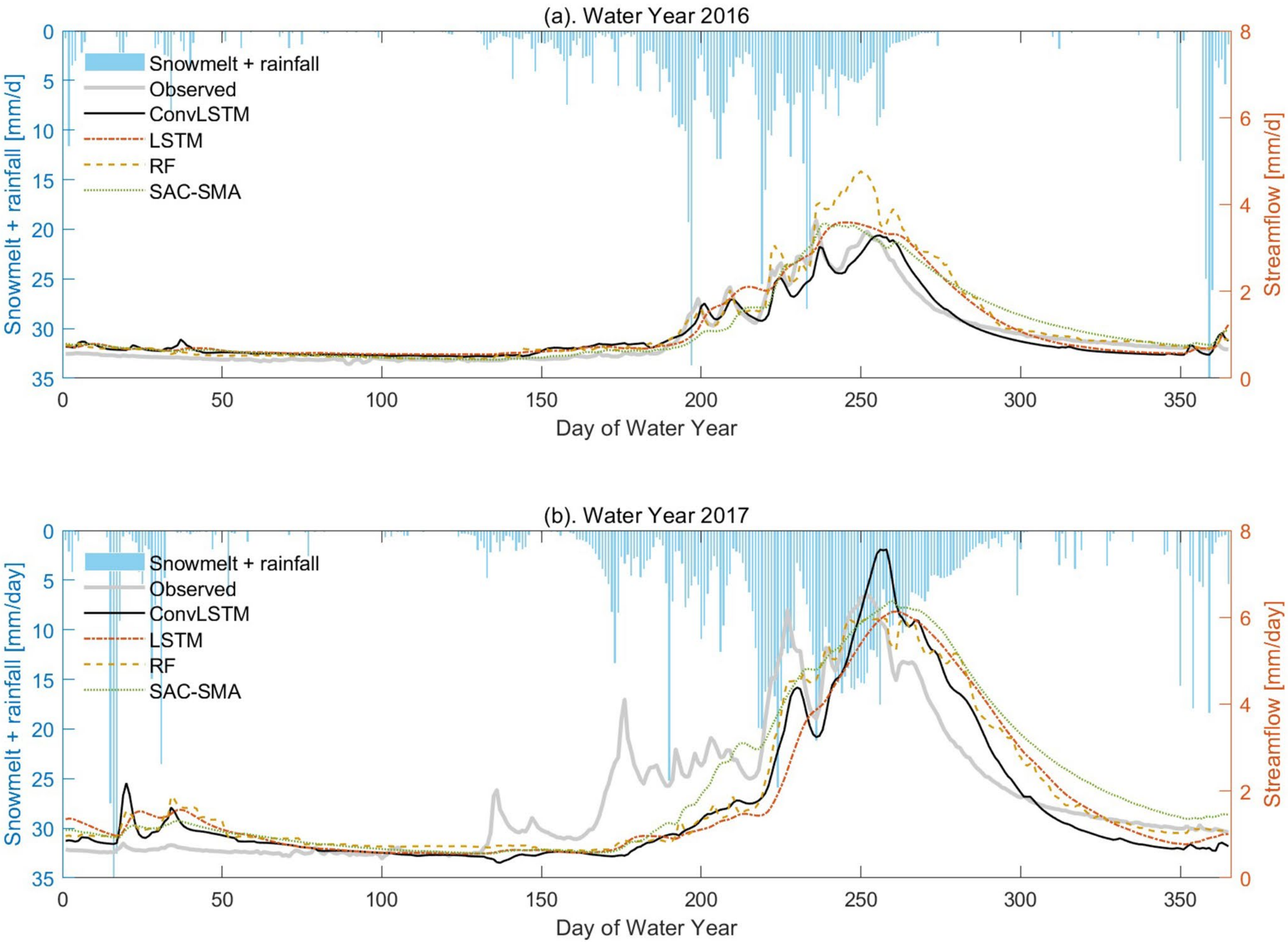

Figure 6. UEB simulated snowmelt and rainfall, observed streamflow, and simulations of ConvLSTM and three reference models for (a) a normal year (WY 2016) and (b) a wet year (WY 2017).

ConvLSTM model yielded similar training and test errors, suggesting drop out and L_2 regularization prevented the complicated model from overfitting. The LSTM, SAC-SMA and RF models show increasing degree of overfitting, as suggested by the larger difference between training and test performance metrics. A large negative PBIAS (overestimation) was found in predictions of the RF and SAC-SMA models.

Precipitation measurements at SNOTEL stations, subject to undercatch, were used to bias correct NLDAS precipitation. The underestimation of precipitation may propagate to the simulated snowmelt rates. Therefore, an "ideal" model is expected to yield a positive PBIAS in streamflow. The study area saw a declining trend of streamflow during the study period (slope = −2.9 mm/year, $p = 0.2$, Figure S3 in Supporting Information S1). Although the decline in the linear trend is not statistically significant, the average annual streamflow during the test period is 5% lower than that of the training period. A similar trend was observed in streamflow simulated by ConvLSTM. In contrast, SAC-SMA simulated streamflow showed an increasing trend (4.5 mm/year), which partially explains the small calibration PBIAS and negative test PBIAS (overestimation). On the other hand, Figure 5 shows the machine learning models often overestimate winter low flows in dry years.

4.3. Spatiotemporal Streamflow Responses to Snowmelt and Rainfall

The streamflow difference (ΔQ) simulated by the ConvLSTM and reference models show significantly different responses to snowmelt pulses with spatially uniform rates (20 mm for one day, and 5 mm for 4 days) (Figure 7). Overall, the streamflow recessions after snowmelt pulses simulated by the machine learning models are faster than the SAC-SMA model. The ConvLSTM model simulation shows two peaks, occurring at similar times as

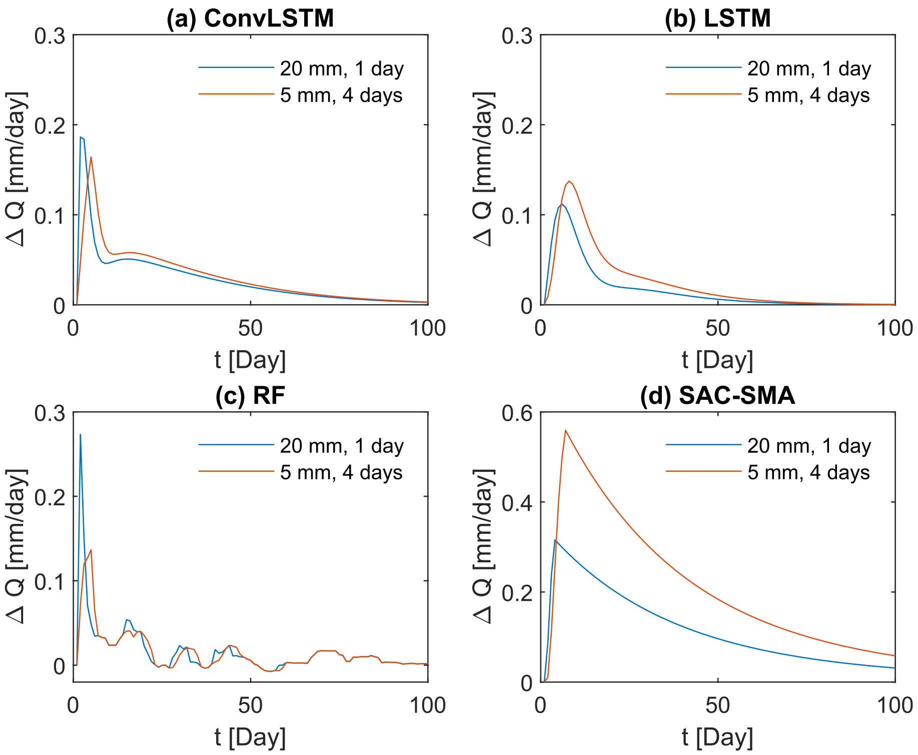

Figure 7. Time varying streamflow change (ΔQ) induced by snowmelt pulses (blue: 20 mm for 1 day, red: 5 mm for 4 days) simulated by (a) Convolutional Long Short-Term Memory, (b) LSTM, (c) RF, and (d) SAC-SMA models. The pulses were introduced beginning at $t = 0$. Note that the scale of the vertical axis in panel (d) is different from other panels.

the first two peaks simulated by RF. On the other hand, the RF model produced fluctuating streamflow response, likely an artifact of the lagged input features (Table S2 in Supporting Information S1). The streamflow differences responding to the 20 mm, 1 day snowmelt pulse simulated by the LSTM and SAC-SMA models are lower than ΔQ responding to the 5 mm, 4 days pulse (Figures 7b and 7d). The streamflow difference simulated by the SAC-SMA model is substantially higher than the other models.

The spatial sensitivities calculated by applying weekly perturbation to daily snowmelt and rainfall rates show variability both in space and time (Figure 8). Spatial sensitivities for perturbations 3 weeks ago were similar to the two-week-ago results, while purturbations more than 3 weeks ago led to lower sensivity (Figure 7). Sensitivity is higher during spring runoff (27 May 2009) than the recession period in summer (1 August 2009). Spatially, higher sensitivity occurs near the west boundary and eastern part of the study area. The high sensitivity area on the west overlaps with the Logan Peak syncline, which controls the movement of groundwater on the west side of Logan River (Spangler, 2001). The subsurface connectivity revealed by previous tracer studies (Figure 1, Spangler, 2001, 2011) illustrate that the recharge area of the Dewitt Springs is consistent with the high sensitivity area during both spring runoff and recession periods. For Wood Camp and Rick's Springs, the recharge areas overall fall within the high sensitivity area of one-week-ahead perturbation. However, a portion of the areas shows negative sensitivity to perturbation 2 weeks prior. Because discharge rates of the other two springs (Benchmark and Logan Cave) are much smaller than the major springs and the Logan River, the sensitivity analysis may be unable to detect their influence.

5. Discussion

5.1. Performance of the Hybrid Modeling Approach

Results of high-resolution snow modeling show a high degree of spatial variability in melt rate and duration (Figure 3). In most of the study area, snow typically starts to melt in April and melt out in June, while at high elevations and north-facing slopes snowpack can last till next water year (Figure 3c). Methods such as temperature

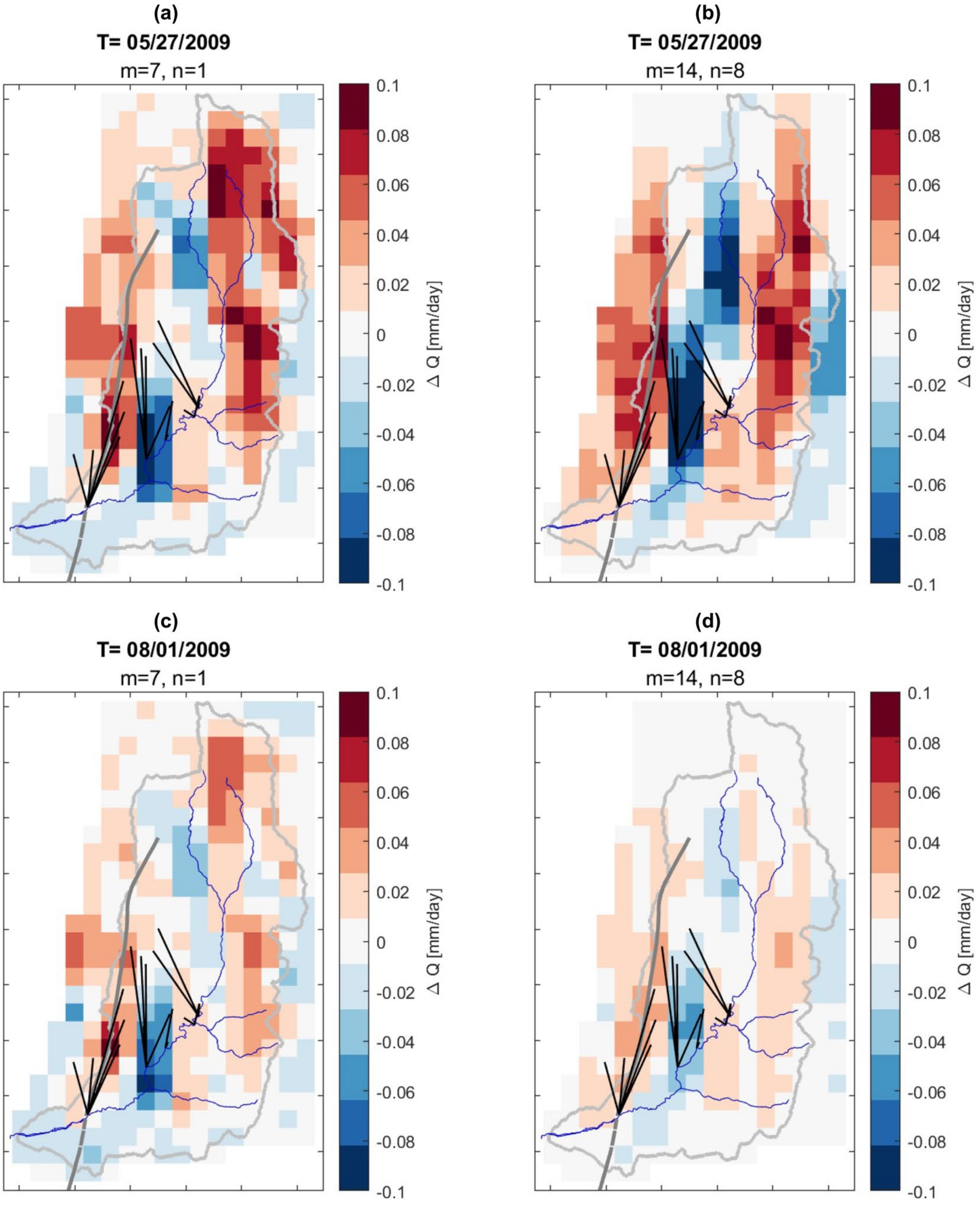

Figure 8. Spatial sensitivity calculated as the change of streamflow on day T (ΔQ_T) when adding weekly perturbation to snowmelt and rainfall on days $T - m, \ldots, T - n$, where T is 27 May 2009 (a–b) and 1 August 2009 (c–d), respectively. Warm color is expected for recharge areas. The dark gray lines along west boundary of the watershed show the axis of the Logan Peak Syncline. Dark lines show subsurface connectivity indicated by tracer studies (Figure 1, Spangler et al., 2001, 2011). The Logan River and tributaries are shown in blue.

index commonly used in lumped and semi-distributed hydrologic models will likely be inadequate to characterize the high degree of spatial variability in complex terrains. It is worth mentioning that since machine learning models use inputs linearly scaled to (0,1), these models utilize "relative" information rather than absolute values. Despite this, the models need accurate representation of spatial and temporal variability in snow processes to accurately simulate streamflow and properly infer recharge-discharge patterns, because a deep snowpack would sustain streamflow later into summer than a more uniform and shallower snowpack. This highlights the importance of process-based modeling of the snow processes at a sufficiently high spatial resolution.

Overall, the ConvLSTM and three reference models all achieved satisfactory performance in simulating streamflow of the study watershed. One exception was found in WY 2017, during which all models yielded later recession than observed (Figure 6). In this year, the melt out dates simulated by UEB were 6–14 days later than observed at SNOTEL stations, likely due to uncertainties in NLDAS forcings.

Among the four models, ConvLSTM performed the best according to the performance metrics (Table 1) during the test period for the study watershed. ConvLSTM did particularly well during the spring runoff period due to its capability to capture streamflow responses to snowmelt/rainfall events (Figures 5 and 6). SAC-SMA also yielded high performance metrics, highlighting the capability of its design to represent a variety of recession behaviors found in nature. It is noteworthy that the SAC-SMA simulated streamflow is less sensitive to snowmelt events than the machine learning models and shows a slower recession (Figures 5 and 6).

The three machine learning models overestimated winter low flows in dry years (Figure 5). Observed low flow exhibits interannual variability (Figure 5, Figure S4 in Supporting Information S1). However, low flows simulated by the machine learning models, especially ConvLSTM and LSTM, are relatively stable. Low flow (typically occurring in February) is correlated with peak SWE in the preceding two calendar years (Figure S4 Supporting Information S1). Although historical snowmelt information is contained in the 365-day lookback and initial cell states, the information appears to be overwritten by more recent inputs due to the long lag (7–8 months) between snowmelt and low flow. Watershed storage memory effects create a longer term dependency than what is commonly seen in conventional LSTM applications such as natural language processing. The results suggest that the LSTM architecture can capture the short term (e.g., within 1 or 2 months) recharge-discharge dynamics, but it may fall short on longer term dynamics in mesoscale watersheds like the Logan River. Recent advances in attention mechanisms that assign higher weights to important input elements (Qin et al., 2017; Vaswani et al., 2017), as well as multi-scale variants of LSTM (Gauch et al., 2021), may have the potential to enhance the capability of LSTMs to maintain information over a longer history. For example, multiple layers of ConvLSTMs or LSTMs can work in parallel to capture watershed dynamics at daily, seasonal, and annual scales. This will enable simulating long-term watershed dynamics, which will help in understanding the resilience of snow-dominated mountainous karst systems under climate variability. Compared to ConvLSTM and LSTM, the RF model performed slightly better during low flow periods, likely because the manually selected features enable the model to partially learn the relation between low flow and historical inputs. The SAC-SMA model captured most of the low flow interannual variability, although possibly due to wrong reasons as explained in Section 5.3.

5.2. Effects of Model Complexity and Sample Size on Predictive Capability

This study included four models with various structures and levels of complexity. For both ConvLSTM and LSTM models, the number of learnable parameters is greater than the sample size (number of training data points). The seeming "overparameterization" is common in deep learning applications and contributes to their high representation power. Concern of overfitting can be alleviated by employing regularization techniques. As a non-parametric regression technique, RF fits a piecewise constant function. Unlike the LSTM and ConvLSTM models that completely rely on the training process to learn their representation of the watershed dynamics, the RF algorithm is less effective with high dimensional inputs and requires domain knowledge for feature engineering. In contrast, the SAC-SMA model structure embeds the conceptualization of key physical processes. With 19 calibrated parameters, the SAC-SMA model is much more parsimonious than the machine learning models. In this study, despite being able to reproduce the hydrograph well, the SAC-SMA model conceptualization that includes representing surface runoff and baseflow recession using separate reservoirs lacks sound hydrologic basis (Klemeš, 1986) especially for this karst watershed. Because SAC-SMA cannot fully resolve the dynamics of the karst watershed, the calibrated parameters may have compensated for model structural error (Doherty & Welter, 2010). Parameters may have also compensated for errors in UEB simulated snowmelt due to precipitation undercatch at SNOTEL stations. Parameter compensation can lead to satisfactory test performance when the hydrologic conditions during the test period are similar to the conditions in the calibration period, but have deleterious implications otherwise (Doherty & Christensen, 2011). In contrast, highly parameterized models tend to be less prone to the deleterious impacts (Doherty & Welter, 2010). In addition, the structure of conceptual hydrologic models such as SAC-SMA may result in low degrees of freedom and thus limit the transferability of a calibrated model to other regions (Nearing et al., 2021).

Previous studies, based on empirical evidence in other watersheds (Ayzel & Heistermann, 2021; Boughton, 2007; Yapo et al., 1996), suggested that the performance of conceptual rainfall-runoff models and deep learning models such as LSTM tend to reach a plateau when the calibration data is longer than a certain duration (e.g., 8–15 years, depending on watershed attributes and climatic conditions). While a comprehensive investigation of sample size effects is beyond the scope of the study, our preliminary analyses suggested a similar performance plateau for the RF model. The performance plateau may indicate the maximum information content these models can get from the data. This limit is affected by the performance criterion used for calibration/training (e.g., MSE, NSE, KGE, or likelihood) (Gupta et al., 2009; Yapo et al., 1996) as well as the representation power of the model structure. Because of the added complexity and capability to digest spatially distributed snowmelt, the ConvLSTM model can flexibly extract information of streamflow responses from streamflow records in an inductive, data-driven way that is not bound by assumptions embedded in conceptual or physically based hydrologic models. Therefore, it may benefit more from longer training periods than other models.

5.3. Is the Streamflow Response Learned by the Models Physically Sensible?

Amid the success of deep learning in terms of high predictive accuracy, there are emerging concerns regarding spurious patterns being learned due to confounding rather than causal relationships (Kaushik et al., 2020). Confounding effects were found in the two-input ConvLSTM and LSTM models. Specifically, the trained models learned a positive correlation between streamflow and PET. This is not surprising because increases in temperature lead to higher snowmelt, thus increasing streamflow. Meanwhile, warmer temperature means higher PET rates. While SAC-SMA "knows" that an increase in PET will produce a non-positive change in streamflow, machine learning models lack such physical knowledge. In this study, the two-output ConvLSTM and LSTM models were unable to distinguish the spurious correlation between streamflow and PET, due to temporal coincidence, from the causal relationship between streamflow and snowmelt. The confounding effect underlines the need for physics informed constraints to prevent deep learning models from picking up such patterns. Meanwhile, RF with feature engineering is partially immune to the confounding effects. It learned negative correlation between streamflow with accumulated PET over the last 150 days, but positive correlation with PET in recent days (Figure S2 in Supporting Information S1).

While all models yielded similar streamflow simulation results, how they arrived at the result is quite different. The first peak in streamflow simulated by ConvLSTM and RF occurs shortly after the snowmelt events ended (Figures 7a and 7c) and may be related to "old" water pushed out of karst conduits and/or both infiltration excess (frozen soil, Niu and Yang, 2006) and saturation excess (during peak snowmelt, Kampf et al., 2015) overland flow. The second, lower peak occurs within the travel time suggested by previous tracer studies (Spangler, 2001, 2011) and thus likely corresponds to snowmelt event water arriving at the Logan River channel through karst conduits. In addition, overland flow and interflow (shallow flow through soil) (Carroll et al., 2019) may be intercepted by sinkholes and losing streams to enter karst conduits. According to the LSTM and SAC-SMA models, the ΔQ induced by a less intense but longer snowmelt event is higher than a more intense but shorter event. This is contradictory to the expectation that intense snowmelt/rainfall events would lead to flashy streamflow surge. The calibrated SAC-SMA model has relatively high depletion rates of lower zone primary and supplemental free water storages and a slow recession unit hydrograph (Table S3 in Supporting Information S1). As a result, the baseflow simulated by the SAC-SMA responds fast to snowmelt and rainfall. Because the tension water storage has been filled up before the pulse, fast baseflow recession produces higher streamflow peaks than simulated by the machine learning models. In contrast to fast baseflow recession, surface runoff and interflow are routed using the unit hydrograph, leading to an almost completely dissipated peak occurring more than 600 days after the snowmelt pulse. Because a larger portion of the 20 mm-intense snowmelt pulse becomes surface runoff, SAC-SMA simulated streamflow responding to 20 mm-intense snowmelt is lower than in the case of 5 mm-intense snowmelt (Figure 8d). Similar phenomena, although to a lesser degree, was observed in the streamflow response simulated by the LSTM model. The differences in ConvLSTM and LSTM results suggest the value of representing spatial variability in snowmelt.

The streamflow recession after snowmelt pulses simulated by the machine learning models is faster than the SAC-SMA model (Figure 7). The fast recession simulated by the machine learning models likely caused the inability to capture the interannual variability of low flow (Figure S4 in Supporting Information S1). As discussed in Section 5.1, the underlying reason may be the limitation of LSTM in representing long-term memory effects governed by watershed storage.

Examining the hydrologic behavior learned by the models shows the seasonal differences in sensitivities (Figure 8) that are consistent with the understanding that when watershed storage is at a higher level, a larger portion of rainfall and snowmelt will produce streamflow. In addition, the spatial patterns of sensitivity can be partially explained by the hydrogeology of the study area. The high sensitivity area on the west coincides with the high elevation portion of the outcrop of the Logan Peak syncline. In this area, sinkholes and faults developed in carbonate units form pathways for fast recharge and conduit flow. Noteworthy, positive sensitivity was found in Water Canyon and Green Canyon (Figure 1), where tracer studies suggested karst piracy (Spangler, 2001, 2011). Such allogenic recharge (Hartmann et al., 2014) and transboundary flow would be challenging to capture using hydrologic models that rely on topographically delineated watershed boundaries. This also highlights the potential of the hybrid modeling approach to extend to other mountainous karst watersheds because it does not require prior information regarding transboundary karst flow. In addition to the major springs where the tracer studies have been performed, there are several springs on both sides of the Logan River near the Klondike Narrows SNOTEL station (Figure 1). While the recharge areas of these springs have not been mapped, they are likely consistent with the high sensitivity areas near the north end of the Logan River Syncline and north to the confluence with Beaver Creek (Figure 1). On the east of the Logan River, the east limb of the syncline outcrops, and the carbonate units are typically overlain by the Wasatch Formation, which can be highly faulted (Dover, 1995; Spangler, 2011). There are several known sinks and faults in this area (Figure 1; Kolesar et al., 2005), likely creating concentrated recharge to the karst aquifer. For example, the Temple Fork subwatershed has several karst springs that may receive recharge from sinkholes in this area (Figure 1). On the other hand, some inconsistencies between the high sensitivity areas and a portion of the recharge areas of Wood Camp and Rick's Springs delineated by tracer studies (Spangler, 2001, 2011) are noticeable. The negative sensitivities found in this area (Figure 8) is counterintuitive because an increase in snowmelt and rainfall is usually expected to increase streamflow. The negative sensitivities may be related to losing condition along the Logan River reach between Wood Camp and Dewitt Springs, but could also suggest physical inconsistency of learned streamflow response to local snowmelt and rainfall. Despite this, the results overall show the potential for the hybrid modeling approach to infer spatial rainfall-discharge patterns from streamflow at the watershed outlet, which is the convolution of local recharge-discharge responses. The local responses are added up differently each year, because snowmelt and rainfall vary spatially and temporally, as captured by the UEB model. Therefore, the multi-year outlet streamflow contains information that enables "deconvolution" of the spatial recharge-discharge patterns by the deep learning algorithm. When available, longer training period and discharge measurements at the subwatershed scale may provide additional information and thus improve the physical consistency of the learned spatial patterns. Meanwhile, the identification of high sensitivity area by the ConvLSTM model can guide the design of future tracer studies to delineate recharge areas on the northern and eastern part of the watershed, which have not been studied yet.

Other studies have suggested connection between watershed storage dynamics and the input-to-state and state-to-state transitions represented by LSTM and other RNNs (Jiang et al., 2020; Kratzert, Herrnegger, et al., 2019). In the context of ConvLSTM, the future state (related to water storage) at a grid is determined by the inputs (inflow) and current states at this grid and its neighbors. Snowmelt and rainfall falling on one grid may recharge and alter the local water storage and/or flow to a nearby grid via surface runoff. This is represented by the convolutional input-to-state transition. Meanwhile, convolutional state-to-state transitions correspond to the distribution of the water storage within the neighborhood via subsurface flow. The size of the neighborhood is determined by the kernel size and time step. Given a fixed time step (1 day in this study), a larger input-to-state kernel may be suitable for faster runoff, and a smaller state-to-state kernel represents slower subsurface flow. In this study, we have used 3×3 kernels for all convolutional operations and stacked three ConvLSTM layers. In addition, while convolutional neural networks have achieved great success for image recognition, some of their design features such as spatial invariance may not be appropriate for hydrologic applications. For example, shifting the location of an object in a photo does not affect its label (e.g., "cat" vs. "dog"). However, recharge occurring at different locations will likely affect streamflow differently for karst watersheds. A potential direction of future work is to use process understanding to design deep learning architectures that can better represent temporal and spatial recharge-discharge patterns.

6. Conclusions

A hybrid modeling approach is presented for simulating streamflow in a snow dominated mountainous karst watershed. The approach used a high-resolution energy balance snow model to characterize the high spatial variability of snow accumulation and melt controlled by the complex terrain and climate in these watersheds. The simulated snowmelt was then used to drive a deep learning model based on the ConvLSTM architecture that learns streamflow response to spatial and temporally varying snowmelt. The hybrid approach only requires meteorological forcing, topography, land cover, and discharge at the watershed outlet, all of which are available in national datasets. In addition, it does not rely on conceptualization of the rainfall-runoff processes, which are challenging for karst watersheds and site specific. As such, the hybrid models are easy to set up for other mountainous karst watersheds without the need for altering the general model structure.

Based on a case study in the Logan River watershed, the hybrid models achieved satisfactory performance and outperformed three lumped rainfall-runoff models based on LSTM, Random Forest, and SAC-SMA, respectively during the test period. Furthermore, interpretative analyses revealed realistic spatial and temporal recharge-discharge patterns learned by the ConvLSTM model. These patterns were relatively consistent with findings from previous hydrogeologic studies in this area. Our results suggest the potential for the hybrid modeling approach for simulating streamflow from snow dominated mountainous karst watersheds, particularly when there are needs for (a) higher accuracy than can be achieved by lumped models, and (b) representation of spatial patterns when detailed subsurface information is not available to support physically based distributed karst modeling.

Lastly, we highlight several points that need further investigation. First, the performance of the hybrid modeling approach for other snow dominated, mountainous, karst watersheds and sample size effects on the performance remain to be quantified. Due to its complexity, the ConvLSTM model may benefit from more training data, possibly from more watersheds, while the performance of conceptual rainfall-runoff models (SAC-SMA) and conventional machine learning algorithms (Random Forest) may have reached a plateau. Second, there are questions regarding how the incorporation of physical knowledge of watershed dynamics can be used to guide the configuration of deep learning architectures and improve the physical consistency of learned recharge-discharge patterns. For example, spatial and temporal attention mechanisms may be a promising alternative to the ConvLSTM architecture to better represent spatially and temporally varying recharge-discharge patterns. Third, the learned recharge-discharge patterns can be further tested using field data (e.g., ion and isotope) in order to generate new understanding of the dynamics of snow dominated, mountainous, karst watersheds.

Data Availability Statement

All data used in this research are publicly available. The UEB software is available at https://hydrology.usu.edu/dtarb/snow/snow.html. The data and code used for simulating streamflow are available at https://github.com/pseudoszechwaniens/UEB_ConvLSTM_model/commits/v1.0.0 (https://doi.org/10.5281/zenodo.5719348).

Acknowledgments

This work was supported by the US National Science Foundation Hydrologic Sciences program grants 2044051/2043363, the USGS 104b, the Utah Water Research Laboratory, and the Logan River Observatory. TX is also supported by NSF grant OAC-1931297. RZ is supported by NOAA CPO-COM grant NA20OAR4310341. The authors thank Hyrum Tennant for his help with data collection and three anonymous reviewers and the Associate Editor for their thoughtful review and suggestions.

References

Addor, N., Newman, A. J., Mizukami, N., & Clark, M. P. (2017). The CAMELS data set: Catchment attributes and meteorology for large-sample studies. *Hydrology and Earth System Sciences*, *21*(10), 5293–5313. https://doi.org/10.5194/hess%2D21%2D5293%2D2017

Anderson, S., & Radic, V. (2021). Evaluation and interpretation of convolutional-recurrent networks for regional hydrological modelling. *Hydrology and Earth System Sciences Discussions*, 1–43.

Andreo, B., Goldscheider, N., Vadillo, I., Vías, J. M., Neukum, C., & Sinreich, M., et al. (2006). Karst groundwater protection: First application of a Pan-European Approach to vulnerability, hazard and risk mapping in the Sierra de Líbar (Southern Spain). *Science of the Total Environment*, *357*(1–3), pp.54–73. https://doi.org/10.1016/j.scitotenv.2005.05.019

Avanzi, F., De Michele, C., Ghezzi, A., Jommi, C., & Pepe, M. (2014). A processing–modeling routine to use SNOTEL hourly data in snowpack dynamic models. *Advances in Water Resources*, *73*, 16–29. https://doi.org/10.1016/j.advwatres.2014.06.011

Ayzel, G., & Heistermann, M. (2021). The effect of calibration data length on the performance of a conceptual hydrological model versus LSTM and GRU: A case study for six basins from the CAMELS dataset. *Computers & Geosciences*, 104708. https://doi.org/10.1016/j.cageo.2021.104708

Bahr, K. (2016). *Structural and lithological influences on the Tony Grove alpine karst system, Bear River range, north central Utah*. PhD thesis. Utah State University.

Barnett, T. P., Adam, J. C. & Lettenmaier, D. P. (2005). Potential impacts of a warming climate on water availability in snow-dominated regions. *Nature*, *438*(7066), p.303.

Bartelt, P., & Lehning, M. (2002). A physical SNOWPACK model for the Swiss avalanche warning: Part I: Numerical model. *Cold Regions Science and Technology*, *35*(3), 123–145. https://doi.org/10.1016/s0165-232x(02)00074-5

Berghuijs, W. R., Woods, R. A., & Hrachowitz, M. (2014). A precipitation shift from snow towards rain leads to a decrease in streamflow. *Nature Climate Change*, *4*(7), 583–586. https://doi.org/10.1038/nclimate2246

Boughton, W. C. (2007). Effect of data length on rainfall–runoff modelling. *Environmental Modelling & Software*, *22*(3), 406–413. https://doi.org/10.1016/j.envsoft.2006.01.001

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Burnash, R. J., Ferral, R. L., & 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, Department of Water Resources.

Butscher, C., & Huggenberger, P. (2008). Intrinsic vulnerability assessment in karst areas: A numerical modeling approach. *Water Resources Research*, *44*, W03408. https://doi.org/10.1029/2007WR006277

Carroll, R. W., Deems, J. S., Niswonger, R., Schumer, R., & Williams, K. H. (2019). The importance of interflow to groundwater recharge in a snowmelt-dominated headwater basin. *Geophysical Research Letters*, *46*(11), 5899–5908. https://doi.org/10.1029/2019gl082447

Chang, Y., Wu, J., Jiang, G., & Kang, Z. (2017). Identification of the dominant hydrological process and appropriate model structure of a karst catchment through stepwise simplification of a complex conceptual model. *Journal of Hydrology*, *548*, 75–87. https://doi.org/10.1016/j.jhydrol.2017.02.050

Chen, F., Barlage, M., Tewari, M., Rasmussen, R., Jin, J., Lettenmaier, D., et al. (2014). Modeling seasonal snowpack evolution in the complex terrain and forested Colorado headwaters region: A model intercomparison study. *Journal of Geophysical Research: Atmospheres*, *119*(24), 13–795. https://doi.org/10.1002/2014jd022167

Chen, T., Li, M., Li, Y., Lin, M., Wang, N., Wang, M., et al. (2015). *Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems*. *arXiv preprint arXiv:1512.01274*.

Chen, Z., Hartmann, A., Wagener, T., & Goldscheider, N. (2018). Dynamics of water fluxes and storages in an Alpine karst catchment under current and potential future climate conditions. *Hydrology and Earth System Sciences*, *22*(7), 3807–3823. https://doi.org/10.5194/hess-22-3807-2018

Cloke, H. L., & Pappenberger, F. (2008). Evaluating forecasts of extreme events for hydrological applications: An approach for screening unfamiliar performance measures. *Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling*, *15*(1), 181–197. https://doi.org/10.1002/met.58

Coulston, J. W., Moisen, G. G., Wilson, B. T., Finco, M. V., Cohen, W. B., & Brewer, C. K. (2012). Modeling percent tree canopy cover: A pilot study. *Photogrammetric Engineering & Remote Sensing*, *78*(7), 715–727.

De Jong, C., Cappy, S., Finckh, M., & Funk, D. (2008). A transdisciplinary analysis of water problems in the mountainous karst areas of Morocco. *Engineering Geology*, *99*(3-4), pp.228–238. https://doi.org/10.1016/j.enggeo.2007.11.021

Doherty, J., & Christensen, S. (2011). Use of paired simple and complex models to reduce predictive bias and quantify uncertainty. *Water Resources Research*, *47*, W12534. https://doi.org/10.1029/2011WR010763

Demargne, J., Wu, L., Regonda, S. K., Brown, J. D., Lee, H., & He, M. et al. (2014). The science of NOAA's operational hydrologic ensemble forecast service. *Bulletin of the American Meteorological Society*, *95*(1), 79–98. https://doi.org/10.1175/BAMS-D-12-00081.1

Doherty, J., & Welter, D. (2010). A short exploration of structural noise. *Water Resources Research*, *46*, W05525. https://doi.org/10.1029/2009WR008377

Dover, J. H. (1995). *Geologic map of the logan 30' x 60' quadrangle, cache and rich counties, Utah, and lincoln and Uinta counties*. U.S. Geological Survey Miscellaneous Investigations Series Map. I-2210.

El-Hakim, M., & Bakalowicz, M. (2007). Significance and origin of very large regulating power of some karst aquifers in the Middle East. Implication on karst aquifer classification. *Journal of Hydrology*, *333*(2–4), 329–339. https://doi.org/10.1016/j.jhydrol.2006.09.003

Elshorbagy, A., Corzo, G., Srinivasulu, S., & Solomatine, D. P. (2010). Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology-Part 2: Application. *Hydrology and Earth System Sciences*, *14*(10), 1943–1961. https://doi.org/10.5194/hess-14-1943-2010

Etchevers, P., Martin, E., Brown, R., Fierz, C., Lejeune, Y., Bazile, E., et al. (2004). Validation of the energy budget of an alpine snowpack simulated by several snow models (SnowMIP project). *Annals of Glaciology*, *38*, 150–158. https://doi.org/10.3189/172756404781814825

Fang, K., Pan, M., & Shen, C. (2018). The value of SMAP for long-term soil moisture estimation with the help of deep learning. *IEEE Transactions on Geoscience and Remote Sensing*, *57*(4), 2221–2233.

Fleming, S. W., Bourdin, D. R., Campbell, D., Stull, R. B., & Gardner, T. (2015). Development and operational testing of a super-ensemble artificial intelligence flood-forecast model for a Pacific Northwest river. *Journal of the American Water Resources Association*, *51*, 502–512. https://doi.org/10.1111/jawr.12259

Fleming, S. W., Garen, D. C., Goodbody, A. G., McCarthy, C. S., & Landers, L. C. (2021). Assessing the new Natural Resources Conservation Service water supply forecast model for the American West: A challenging test of explainable, automated, ensemble artificial intelligence. *Journal of Hydrology*, *602*, 126782. https://doi.org/10.1016/j.jhydrol.2021.126782

Fleming, S. W., & Goodbody, A. G. (2019). A machine learning metasystem for robust probabilistic nonlinear regression-based forecasting of seasonal water availability in the US West. *IEEE Access*, *7*, 119943–119964. https://doi.org/10.1109/access.2019.2936989

Fleming, S. W., Vesselinov, V. V., & Goodbody, A. G. (2021). Augmenting geophysical interpretation of data-driven operational water supply forecast modeling for a western US river using a hybrid machine learning approach. *Journal of Hydrology*, *597*, 126327. https://doi.org/10.1016/j.jhydrol.2021.126327

Fleury, P., Plagnes, V., & Bakalowicz, M. (2007). Modelling of the functioning of karst aquifers with a reservoir model: Application to Fontaine de Vaucluse (South of France). *Journal of Hydrology*, *345*, 38–49. https://doi.org/10.1016/j.jhydrol.2007.07.014

Gauch, M., Kratzert, F., Klotz, D., Nearing, G., Lin, J., & Hochreiter, S. (2021). Rainfall–runoff prediction at multiple timescales with a single Long Short-Term Memory network. *Hydrology and Earth System Sciences*, *25*(4), 2045–2062. https://doi.org/10.5194/hess-25-2045-2021

Gislason, P. O., Benediktsson, J. A., & Sveinsson, J. R. (2006). Random forests for land cover classification. *Pattern Recognition Letters*, *27*(4), 294–300. https://doi.org/10.1016/j.patrec.2005.08.011

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Groisman, P. Y., & Easterling, D. R. (1994). Variability and trends of total precipitation and snowfall over the United States and Canada. *Journal of Climate*, *7*(1), 184–205. https://doi.org/10.1175/1520-0442(1994)007<0184:vatotp>2.0.co;2

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, *377*(1–2), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Gupta, H. V., Sorooshian, S., & Yapo, P. O. (1999). Status of automatic calibration for hydrologic models: Comparison with multilevel expert calibration. *Journal of Hydrologic Engineering*, *4*(2), 135–143. https://doi.org/10.1061/(asce)1084-0699(1999)4:2(135)

Harpold, A., Brooks, P., Rajagopal, S., Heidbuchel, I., Jardine, A., & Stielstra, C. (2012). Changes in snowpack accumulation and ablation in the intermountain West. *Water Resources Research*, *48*(11). https://doi.org/10.1029/2012wr011949

Hartmann, A., Goldscheider, N., Wagener, T., Lange, J., & Weiler, M. (2014). Karst water resources in a changing world: Review of hydrological modeling approaches. *Reviews of Geophysics*, *52*(3), 218–242. https://doi.org/10.1002/2013rg000443

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning*. Springer series in statistics.

He, X., Chaney, N. W., Schleiss, M., & Sheffield, J. (2016). Spatial downscaling of precipitation using adaptable random forests. *Water Resources Research*, *52*(10), 8217–8237. https://doi.org/10.1002/2016wr019034

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). *Improving neural networks by preventing co-adaptation of feature detectors*. arXiv preprint arXiv:1207.0580.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hood, J. L., & Hayashi, M. (2015). Characterization of snowmelt flux and groundwater storage in an alpine headwater basin. *Journal of Hydrology*, *521*, 482–497. https://doi.org/10.1016/j.jhydrol.2014.12.041

Hsu, K. L., Gao, X., Sorooshian, S., & Gupta, H. V. (1997). Precipitation estimation from remotely sensed information using artificial neural networks. *Journal of Applied Meteorology*, *36*(9), 1176–1190. https://doi.org/10.1175/1520-0450(1997)036<1176:pefrsi>2.0.co;2

Jia, X., Willard, J., Karpatne, A., Read, J., Zwart, J., Steinbach, M., & Kumar, V. (2019). Physics guided RNNs for modeling dynamical systems: A case study in simulating lake temperature profiles. In *Proceedings of the 2019 SIAM international conference on data mining* (pp. 558–566). Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611975673.63

Jiang, S., Zheng, Y., & Solomatine, D. (2020). Improving AI system awareness of geoscience knowledge: Symbiotic integration of physical approaches and deep learning. *Geophysical Research Letters*, *47*(13), e2020GL088229. https://doi.org/10.1029/2020gl088229

Jordan, R. (1991). A one-dimensional temperature model for a snow cover: Technical documentation for SNTHERM. No. CRREL-SR-91-16. *Cold Regions Research and Engineering Lab*, *89*.

Kampf, S., Markus, J., Heath, J., & Moore, C. (2015). Snowmelt runoff and soil moisture dynamics on steep subalpine hillslopes. *Hydrological Processes*, *29*(5), 712–723. https://doi.org/10.1002/hyp.10179

Karpatne, A., Atluri, G., Faghmous, J. H., Steinback, M., Banerjee, A., Ganguly, A., et al. (2017). Theory-guided data science: A new paradigm for scientific discover from data. *IEEE Transactions on Knowledge and Data Engineering*, *29*, 2318–2331. https://doi.org/10.1109/tkde.2017.2720168

Kaushik, D., Hovy, E., & Lipton, Z. C. (2020). Learning the difference that makes a difference with counterfactually-augmented data. In *Proceedings of ICLR*.

Klemeš, V. (1986). Dilettantism in hydrology: Transition or destiny? *Water Resources Research*, *22*(9S), 177S–188S.

Kolesar, P. T., Evans, J. P., Gooseff, M. N., Lachmar, T. E., & Payn, R. (2005). A tale of two (or more) karsts, Bear River range, Cache National Forest. *Utah Geol Soc Am Abstr Programs*, *37*(7), 177.

Kratzert, F., Herrnegger, M., Klotz, D., Hochreiter, S., & Klambauer, G. (2019). NeuralHydrology–interpreting LSTMs in hydrology. In *Explainable AI: Interpreting, explaining and visualizing deep learning* (pp. 347–362). Springer. https://doi.org/10.1007/978-3-030-28954-6_19

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018). Rainfall–runoff modelling using long short-term memory (LSTM) networks. *Hydrology and Earth System Sciences*, *22*(11), 6005–6022. https://doi.org/10.5194/hess-22-6005-2018

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., & Nearing, G. S. (2019). *Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning* (Vol. 55). Water Resources Research. https://doi.org/10.1029/2019wr026065

Lary, D. J., Alavi, A. H., Gandomi, A. H., & Walker, A. L. (2016). Machine learning in geosciences and remote sensing. *Geoscience Frontiers*, *7*(1), 3–10. https://doi.org/10.1016/j.gsf.2015.07.003

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Li, D., Wrzesien, M. L., Durand, M., Adam, J., & Lettenmaier, D. P. (2017). How much runoff originates as snow in the western United States, and how will that change in the future? *Geophysical Research Letters*.

Li, Z., Xu, X., Liu, M., Li, X., Zhang, R., Wang, K., & Xu, C. (2017). State-space prediction of spring discharge in a karst catchment in southwest China. *Journal of Hydrology*, *549*, 264–276. https://doi.org/10.1016/j.jhydrol.2017.04.001

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors*, *18*(8), 2674. https://doi.org/10.3390/s18082674

Liston, G. E., & Elder, K. (2006). A meteorological distribution system for high-resolution terrestrial modeling (MicroMet). *Journal of Hydrometeorology*, *7*(2), 217–234. https://doi.org/10.1175/jhm486.1

Lv, N., Liang, X., Chen, C., Zhou, Y., Li, J., Wei, H., & Wang, H. (2020). A long short-term memory cyclic model with mutual information for hydrology forecasting: A case study in the xixian basin. *Advances in Water Resources*, 103622. https://doi.org/10.1016/j.advwatres.2020.103622

Mahat, V., & Tarboton, D. G. (2012). Canopy radiation transmission for an energy balance snowmelt model. *Water Resources Research*, *48*(1). https://doi.org/10.1029/2011wr010438

Mahat, V., & Tarboton, D. G. (2014). Representation of canopy snow interception, unloading and melt in a parsimonious snowmelt model. *Hydrological Processes*, *28*(26), 6320–6336. https://doi.org/10.1002/hyp.10116

Malard, A., Sinreich, M., & Jeannin, P. Y. (2016). A novel approach for estimating karst groundwater recharge in mountainous regions and its application in Switzerland. *Hydrological Processes*, *30*(13), 2153–2166. https://doi.org/10.1002/hyp.10765

Mazzilli, N., Guinot, V., Jourde, H., Lecoq, N., Labat, D., Arfib, B., et al. (2017). *KarstMod: A modelling platform for rainfall-discharge analysis and modelling dedicated to karst systems*. Environmental Modelling & Software. https://doi.org/10.1016/j.envsoft.2017.03.015

McGovern, A., Elmore, K. L., Gagne, D. J., Haupt, S. E., Karstens, C. D., Lagerquist, R., et al. (2017). Using artificial intelligence to improve real-time decision-making for high-impact weather. *Bulletin of the American Meteorological Society*, *98*(10), 2073–2090. https://doi.org/10.1175/bams-d-16-0123.1

Meyer, J. D., Jin, J., & Wang, S. Y. (2012). Systematic patterns of the inconsistency between snow water equivalent and accumulated precipitation as reported by the snowpack telemetry network. *Journal of Hydrometeorology*, *13*(6), 1970–1976. https://doi.org/10.1175/JHM-D-12-066.1

Mo, S., Zhu, Y., Zabaras, N., Shi, X., & Wu, J. (2019). Deep convolutional encoder-decoder networks for uncertainty quantification of dynamic multiphase flow in heterogeneous media. *Water Resources Research*, *55*(1), 703–728. https://doi.org/10.1029/2018wr023528

Naghibi, S. A., Pourghasemi, H. R., & Dixon, B. (2016). GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. *Environmental monitoring and assessment*, *188*(1), 1–27.

Nash, J. E. (1957). *The Form of the Instantaneous Unit Hydrograph* (Vol. 45, pp. 114–121). International Association of Scientific Hydrology Publication.

Naz, B. S., Kao, S. C., Ashfaq, M., Rastogi, D., Mei, R., & Bowling, L. C. (2016). Regional hydrologic response to climate change in the conterminous United States using high-resolution hydroclimate simulations. *Global and Planetary Change*, *143*, 100–117. https://doi.org/10.1016/j.gloplacha.2016.06.003

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., et al. (2021). What role does hydrological science play in the age of machine learning? *Water Resources Research*, *57*(3), e2020WR028091. https://doi.org/10.1029/2020wr028091

Neilson, B. T., Tennant, H., Stout, T. L., Miller, M., Gabor, R. S., Jameel, Y., et al. (2018). Stream-centric methods for determining groundwater contributions in karst mountain watersheds. *Water Resources Research*, *54*(9), 6708–6724. https://doi.org/10.1029/2018wr022664

Nelson, K. J., Connot, J., Peterson, B., & Martin, C. (2013). The landfire refresh strategy: Updating the national dataset. *Fire Ecol*, *9*, 80–101. https://doi.org/10.4996/fireecology.0902080

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., et al. (2015). Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: Data set characteristics and assessment of regional variability in hydrologic model performance. *Hydrology and Earth System Sciences*, *19*(1), 209–223. https://doi.org/10.5194/hess-19-209-2015

Niu, G. Y., & Yang, Z. L. (2006). Effects of frozen soil on snowmelt runoff and soil water storage at a continental scale. *Journal of Hydrometeorology*, *7*(5), 937–952. https://doi.org/10.1175/jhm538.1

Pan, B., Hsu, K., AghaKouchak, A., & Sorooshian, S. (2019). Improving precipitation estimation using convolutional neural network. *Water Resources Research*, *55*, 2301–2321. https://doi.org/10.1029/2018wr024090

Priestley, C. H. B., & Taylor, R. J. (1972). On the assessment of surface heat flux and evaporation using large-scale parameters. *Monthly Weather Review*, *100*(2), 81–92. https://doi.org/10.1175/1520-0493(1972)100<0081:otaosh>2.3.co;2

Qin, Y., Song, D., Chen, H., Cheng, W., Jiang, G., & Cottrell, G. (2017). *A dual-stage attention-based recurrent neural network for time series prediction*. arXiv preprint arXiv:1704.02971.

Rasouli, K., Hsieh, W. W., & Cannon, A. J. (2012). Daily streamflow forecasting by machine learning methods with weather and climate inputs. *Journal of Hydrology*, *414*, 284–293. https://doi.org/10.1016/j.jhydrol.2011.10.039

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., & Carvalhais, N. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), 195–204. https://doi.org/10.1038/s41586-019-0912-1

Rutter, N., Essery, R., Pomeroy, J., Altimir, N., Andreadis, K., Baker, I., et al. (2009). Evaluation of forest snow processes models (SnowMIP2). *Journal of Geophysical Research*, *114*(D6). https://doi.org/10.1029/2008jd011063

Scanlon, B. R., Mace, R. E., Barrett, M. E., & Smith, B. (2003). Can we simulate regional groundwater flow in a karst system using equivalent porous media models? Case study, barton Springs Edwards aquifer, USA. *Journal of Hydrology*, *276*(1–4), 137–158. https://doi.org/10.1016/s0022-1694(03)00064-7

Schlögl, S., Marty, C., Bavay, M., & Lehning, M. (2016). Sensitivity of Alpine3D modeled snow cover to modifications in DEM resolution, station coverage and meteorological input quantities. *Environmental Modelling & Software*, *83*, 387–396.

Sen Gupta, A., & Tarboton, D. G. (2016). A tool for downscaling weather data from large-grid reanalysis products to finer spatial scales for distributed hydrological applications. *Environmental Modelling & Software*, *84*, 50–69. https://doi.org/10.1016/j.envsoft.2016.06.014

Shen, C., Laloy, E., Elshorbagy, A., Albert, A., Bales, J., Chang, F. J., Fang, K. (2018). HESS Opinions: Incubating deep-learning-powered hydrologic science advances as a community. *Hydrology and Earth System Sciences*, *22*(11). https://doi.org/10.5194/hess-22-5639-2018

Shi, X., Chen, Z., Wang, H., Yeung, D. Y., Wong, W. K., & Woo, W. C. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems* (pp. 802–810).

Solomatine, D. P., & Ostfeld, A. (2008). Data-driven modelling: Some past experiences and new approaches. *Journal of Hydroinformatics*, *10*(1), 3–22. https://doi.org/10.2166/hydro.2008.015

Spangler, L. E. (2001). *Delineation of recharge areas for karst springs in logan canyon, Bear River range, northern Utah* (pp. 01–4011). United States Geological Survey Karst Interest Group Proceedings, Water-Resources Investigations Report.

Spangler, L. E. (2011). *Karst hydrogeology of the Bear River range in the vicinity of the Logan River, northern Utah*. Geological Survey. Paper presented at Geological Society of America Rocky Mountain - Cordilleran section meeting, U.S.

Sturm, M., Goldstein, M. A., & Parr, C. (2017). Water and life from snow: A trillion dollar science question. *Water Resources Research*, *53*(5), 3534–3544. https://doi.org/10.1002/2017wr020840

Sultana, R., Hsu, K. L., Li, J., & Sorooshian, S. (2014). Evaluating the Utah Energy Balance (UEB) snow model in the Noah land-surface model. *Hydrology and Earth System Sciences*, *18*, 3553. https://doi.org/10.5194/hess-18-3553-2014

Sun, A. Y., Scanlon, B. R., Zhang, Z., Walling, D., Bhanja, S. N., Mukherjee, A., & Zhong, Z. (2019). Combining physically based modeling and deep learning for fusing GRACE satellite data: Can we learn from mismatch? *Water Resources Research*, *55*. https://doi.org/10.1029/2018wr023333

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random forest: A classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Sciences*, *43*(6), 1947–1958. https://doi.org/10.1021/ci034160g

Sweeting, M. M. (2012). *Karst in China: Its geomorphology and environment* (Vol. 15). Springer Science & Business Media.

Tarboton, D. G., & Luce, C. H. (1996). *Utah energy balance snow accumulation and melt model (UEB)*. Utah Water Research Laboratory.

Taylor, C. J., & Greene, E. A. (2008). Hydrogeologic characterization and methods used in the investigation of karst hydrology. In D. O. Rosenberry, & J. W. LaBaugh (Eds.), *Field techniques for estimating water fluxes between surface water and ground water* (pp. 71–114). US Geological Survey.

Tennant, C., Larsen, L., Bellugi, D., Moges, E., Zhang, L., & Ma, H. (2020). The utility of information flow in formulating discharge forecast models: A case study from an arid snow-dominated catchment. *Water Resources Research*, *56*(8), e2019WR024908. https://doi.org/10.1029/2019wr024908

Tyson, C. (2021). Effects of Climate Forcing Uncertainty on High-Resolution Snow Modeling and Streamflow Prediction in a Mountainous Karst Watershed. Master's Thesis, Utah State University, Retrived from https://digitalcommons.usu.edu/etd/8041

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).

Vrugt, J. A., Ter Braak, C. J. F., Diks, C. G. H., Robinson, B. A., Hyman, J. M., & Higdon, D. (2009). Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling. *International Journal of Nonlinear Science*, *10*, 273–290. https://doi.org/10.1515/ijnsns.2009.10.3.273

Xia, Y., Mitchell, K., Ek, M., Sheffield, J., Cosgrove, B., Wood, E., et al. (2012). Continental-scale water and energy flux analysis and validation for the North American land data assimilation system project phase 2 (NLDAS-2): 1. Intercomparison and application of model products. *Journal of Geophysical Research*, *117*(D3). https://doi.org/10.1029/2011jd016048

Xiang, Z., Yan, J., & Demir, I. (2020). A rainfall-runoff model with LSTM-based sequence-to-sequence learning. *Water Resources Research*, *56*(1), e2019WR025326. https://doi.org/10.1029/2019wr025326

Xu, T., & Liang, F. (2021). *Machine learning for hydrologic sciences: An introductory overview*. Wiley Interdisciplinary Reviews.e1533

Xu, T., Valocchi, A. J., Ye, M., & Liang, F. (2017). Quantifying model structural error: Efficient Bayesian calibration of a regional groundwater flow model using surrogates and a data-driven error model. *Water Resources Research*, *53*(5), 4084–4105. https://doi.org/10.1002/2016wr019831

Yang, L., Jin, S., Danielson, P., Homer, C. G., Gass, L., Bender, S. M., et al., (2018). A new generation of the United States national land cover database—requirements, research priorities, design, and implementation strategies: ISPRS journal of photogrammetry and remote sensing, v. *146*, p. 108–123, https://doi.org/10.1016/j.isprsjprs.2018.09.006

Yapo, P. O., Gupta, H. V., & Sorooshian, S. (1996). Automatic calibration of conceptual rainfall-runoff models: Sensitivity to calibration data. *Journal of Hydrology*, *181*(1–4), 23–48. https://doi.org/10.1016/0022-1694(95)02918-4

References From the Supporting Information

Hall, D. K., & Riggs, G. A. (2016). MODIS/Terra snow cover daily L3 global 500m SIN grid (Boulder CO, USA: NASA snow and ice data center). Accessed at Aug. 15, 2020.

Yang, J., Jiang, L., Ménard, C. B., Luojus, K., Lemmetyinen, J., & Pulliainen, J. (2015). Evaluation of snow products over the Tibetan Plateau. *Hydrological Processes*, *29*(15), 3247–3260. https://doi.org/10.1002/hyp.10427