Water Resources Research

RESEARCH ARTICLE
10.1029/2021WR031033

Exploring the Potential of Long Short-Term Memory Networks for Improving Understanding of Continental- and Regional-Scale Snowpack Dynamics

Yuan-Heng Wang[1], Hoshin V. Gupta[1], Xubin Zeng[1], and Guo-Yue Niu[1]

[1]Department of Hydrology and Atmospheric Sciences, The University of Arizona, Tucson, AZ, USA

Key Points:

- A trained continental-scale Long Short-Term Memory (LSTM) network is capable of providing almost as good performance as a regionally trained one
- The continent-scale LSTM outperforms a regionally trained SNOW17, and a SNOW17 model calibrated locally to each pixel across the domain
- The LSTM exhibits better spatial transferability than SNOW17, and exhibits a trade-off between transferability and model complexity

Supporting Information:

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

Correspondence to:

Y.-H. Wang,
yhwang0730@email.arizona.edu

Citation:

Wang, Y.-H., Gupta, H. V., Zeng, X., & Niu, G.-Y. (2022). Exploring the potential of long short-term memory networks for improving understanding of continental- and regional-scale snowpack dynamics. *Water Resources Research*, *58*, e2021WR031033. https://doi.org/10.1029/2021WR031033

Received 11 AUG 2021
Accepted 21 FEB 2022

Author Contributions:

Conceptualization: Hoshin V. Gupta
Data curation: Yuan-Heng Wang, Xubin Zeng
Formal analysis: Yuan-Heng Wang
Funding acquisition: Guo-Yue Niu
Investigation: Yuan-Heng Wang, Hoshin V. Gupta
Methodology: Yuan-Heng Wang, Hoshin V. Gupta
Project Administration: Xubin Zeng, Guo-Yue Niu
Resources: Hoshin V. Gupta, Guo-Yue Niu
Software: Yuan-Heng Wang
Supervision: Hoshin V. Gupta, Xubin Zeng

Abstract

Accurate estimation of the spatio-temporal distribution of snow water equivalent is essential given its global importance for understanding climate dynamics and climate change, and as a source of fresh water. Here, we explore the potential of using the Long Short-Term Memory (LSTM) network for continental and regional scale modeling of daily snow accumulation and melt dynamics at 4-km pixel resolution across the conterminous US (CONUS). To reduce training costs (data are available for ∼0.31 million snowy pixels), we combine spatial sampling with stagewise model development, whereby the network is first pretrained across the entire CONUS and then subjected to regional fine-tuning. Accordingly, model evaluation is focused on out-of-sample predictive performance across space (analogous to the prediction in ungauged basins problem). We find that, given identical inputs (precipitation, temperature, and elevation), a single CONUS-wide LSTM provides significantly better spatio-temporal generalization than a regionally calibrated version of the physical-conceptual temperature-index-based SNOW17 model. Adding more meteorological information (dew point temperature, vapor pressure deficit, longwave radiation, and shortwave radiation) further improves model performance, while rendering redundant the local information provided by elevation. Overall, the LSTM exhibits better transferability than SNOW17 to locations that were not included in the training data set, reinforcing the advantages of structure learning over parameter learning. Our results suggest that an LSTM-based approach could be used to develop continental/global-scale systems for modeling snow dynamics.

Plain Language Summary

Understanding the spatio-temporal distribution of water in the snowpack (known as snow water equivalent) is very important for understanding climate dynamics and climate change, and for forecasting and management of global water supplies. In this study, we use Deep Learning (DL) to model snow accumulation and melt at 4-km pixel-scale resolution across the conterminous US (CONUS). Long Short-Term Memory (LSTM) networks are developed at both continental- and regional-scale, by combining spatial pixel sampling and stagewise network pre-training/fine-tuning. We benchmark out-of-sample predictive performance against the physical-conceptual temperature-index-based SNOW17 model, and find that LSTM networks significantly outperform calibrated versions of the SNOW17 model when given identical information. Further, when provided with additional meteorological information, performance of the LSTM is improved. The LSTM models also exhibits better transferability than the SNOW17, indicating the potential for future development of a DL-based system for modeling continental/global-scale snow dynamics.

1. Introduction

1.1. The Problem of Continental-Scale Estimation of Snow Water Equivalent

Accurate monitoring of the large-scale dynamics of snowpack is essential for understanding the details of climate dynamics and climate change (Robinson et al., 1993). Warming under a changing climate is expected to cause snowpack to melt earlier in the year (Xiao, 2021; Zeng et al., 2018) and to reduce the amount of water stored as snow (Musselman et al., 2021; Nijssen et al., 2001). This is expected to have broad and potentially severe impacts to ecosystem productivity (Boisvenue & Running, 2006), winter flood risk (Musselman et al., 2018), groundwater recharge (Ford et al., 2020), agriculture and food security (Qin et al., 2020; Shindell et al., 2012), wildfire hazard (Westerling, 2016), and frequency and severity of drought (Arevalo et al., 2021). In western North America, snow is the primary source of water and streamflow (Li et al., 2017), while globally it supports the water supply needs for more than 1 billion people (Barnett et al., 2005). Therefore, having accurate estimates of the quantity of water stored in snowpack, called snow water equivalent (SWE), is critical for the forecasting and management of water supply and hydropower (Bales et al., 2006; Mankin et al., 2015).

Validation: Yuan-Heng Wang, Hoshin V. Gupta
Visualization: Yuan-Heng Wang
Writing – original draft: Yuan-Heng Wang
Writing – review & editing: Yuan-Heng Wang, Hoshin V. Gupta, Xubin Zeng, Guo-Yue Niu

Several different physically based snow models have been developed to simulate the co-evolution of mass and energy within the snowpack system, and to thereby provide estimates of SWE. Examples include the temperature-index based SNOW-17 model (Anderson, 1973), UEB (Tarboton & Luce, 1996), SAST (Jin et al., 1999), ESCIMO (Strasser et al., 2002), and SNOWCAN (Tribbeck et al., 2004). More sophisticated snow models that focus on advanced representations of stratigraphy or internal dynamics (i.e., grain structure etc.) of the snowpack include Crocus (Brun et al., 1992), and the physics-based SNOWPACK model (Bartelt & Lehning, 2002). In practice, modelers typically use simpler physical-conceptual land-surface representations such as VIC (Liang et al., 1994) to estimate the broad changes in snowpack that might be expected under climate change. Meanwhile, the iSNOBAL model has been the modeling engine for spatially distributed SWE estimation within the Airborne Snow Observatory (ASO) product (Marks et al., 1999).

Nonetheless, the predictive performance of all such models depends on whether or not their representations of the underlying data-generating processes are adequate. To address poor predictive performance stemming from inadequate physical representations, modelers have explored a full spectrum of explicit process hypotheses (Noah-MP; Niu et al., 2011), synthesized multiple working hypotheses into a unifying modeling framework (SUMMA; Clark et al., 2011, 2016), linked the parameter values to local basin attributes by imposing spatial regularization constraints (Pokhrel et al., 2008) via parameter transfer functions (mHM; Samaniego et al., 2010), and explored implementations at finer spatial resolutions (HydroBlocks; Chaney et al., 2016). However, a potential downside of such methods is the large computational demands imposed when conducting simulations at practically useful resolutions over large spatial extents.

Following a complementary approach, statistical data-driven approaches (such as multiple linear regression and binary regression trees) have also been widely used to generate estimates of targeted snow variables at continental- and watershed-scales by exploiting the information provided by field measurements in conjunction with observed physiographic and meteorological covariates (see the review in Broxton et al., 2019). Many studies have explored machine learning (ML) approaches to the estimation of snow variables (e.g., snow depth, snowfall, SWE and the fractional snow cover) include the application of Random Forest and Support Vector Machine methods, using a variety of input data such as satellite sensors (Ehsani, Behrangi, et al., 2021; Kuter, 2021; Kuter et al., 2018), terrestrial laser scanners (Revuelto et al., 2020), land models (Snauffer et al., 2018), and ground observations (Buckingham et al., 2015; Gharaei-Manesh et al., 2016; Tabari et al., 2010). The results of these efforts, which draw upon recent advances in machine learning (ML), and particularly deep learning, suggest that ML-based methods have the potential to outperform state-of-the-art techniques for many sophisticated domain problems (Kratzert, Klotz, Herrnegger, et al., 2019).

In the context of snow hydrology, the artificial neural network (ANN; sometimes called the feedforward multilayer perceptron) has been used to improve the estimation of SWE in different ways, such as the Snow Water Artificial Neural Network Modeling (SWANN) system (Broxton et al., 2017). Snauffer et al. (2018) used ANNs for multi-source data fusion, using SWE data from reanalysis products and manual snow surveys as network inputs, and reported improvements in the quality of gridded SWE products. Broxton et al. (2019) combined aerial remotely sensed maps of snow depth with snow density maps generated via artificial neural network (ANN) processing of field measurements to improve the estimation of SWE. These successes can be attributed to the ability of ANNs to learn the nonlinear nature of the relationships between the relevant variables, resulting in improved performance over traditional statistical methods (Czyzowska-Wisniewski et al., 2015). Recently, studies of how to improve SWE estimates have explored the use of multiple data types and a variety of features derived from meteorological quantities as inputs to the training of ensemble multilayer perceptron (MLP) models (Ntokas et al., 2021; Odry et al., 2020). In general, it seems reasonable that ML-based methods should be able to provide relevant and useful information over large spatial domains; see for example, the pixel scale return-level design maps of SWE developed for modeling snowmelt-driven floods over the entire continental United States, hereafter referred to as CONUS (Cho & Jacobs, 2020; Welty & Zeng, 2021).

1.2. The Potential Offered by Deep Learning

Deep learning (DL) has recently been proposed as a powerful strategy for hydrological modeling and time-series prediction (Shen, 2018; Shen et al., 2018). In particular, the long short-term memory network (LSTM; Hochreiter & Schmidhuber, 1997) has been reported to outperform the traditional ANN approach, provided that sufficient data is available for model development (Wunsch et al., 2021). In particular, Kratzert et al. (2018) showed that the knowledge encapsulated by the generic pre-trained LSTM network can be transferred to different locations in the context of rainfall-runoff modeling. By initializing the LSTM network parameters to those of the pre-trained model, and by conducting subsequent local fine-tuning (Kratzert et al., 2018; Yosinski et al., 2014) it should be possible to reduce local data requirements, thereby facilitating a variety of hydrological applications such as regionalization and prediction in ungauged basins (PUB; Hrachowitz et al., 2013; Sivapalan et al., 2003).

For rainfall-runoff modeling, Kratzert, Klotz, Shalev, et al. (2019) showed that a single regionally trained LSTM network can provide better basin-specific predictions than traditional hydrological models locally calibrated basin-by-basin. Further, when the regionally trained LSTM was applied to basins whose data was not used for model development (i.e., effectively treating them as "ungauged" basins) it performed, on average, better than instances of the Sacramento Soil Moisture Accounting Model (SAC-SMA) or the NOAA National Water Model that were directly calibrated to those same basins (Kratzert, Klotz, Herrnegger, et al., 2019). These asymmetrical comparisons illustrate the ability of a standard LSTM architecture to learn a model structure that performs better than a "physics-based" model, by effectively exploiting the relevant information available in the input-output data.

1.3. Problem Definition, Objectives, and Scope of This Work

This study explores the capability of LSTMs for modeling the dynamics of snow accumulation and melt. The main goal is to achieve accurate estimates of SWE over a large spatial domain by exploiting available pixel-scale datasets while maintaining a reasonable level of computational cost. Our approach involves step-wise training (Kratzert et al., 2018) of an LSTM network using a subset of pixel-scale data sampled across the entire CONUS, where we first use CONUS-wide network pre-training to initialize the network parameters, followed by regional fine-tuning of the network. In particular, our modeling experiments were designed to examine the spatial transferability of predictive performance, thereby facilitating the application of PUB in the context of snow hydrology (Kratzert, Klotz, Herrnegger, et al., 2019).

To explore the best achievable performance for SWE modeling, we train the LSTM networks using different combinations of "available" input data, and benchmark the network performance against the temperature-index-based SNOW17 model (Anderson, 2006; hereafter SN17). Our main interests are in (a) whether the LSTM can outperform the SN17 model used by the National Weather Service River Forecast Center (NWS RFC) for operational hydrologic prediction, and (b) to what extent the performance of the LSTM is affected by different system structure hypotheses, implemented as continental, regional and local training (calibration) strategies.

The scope of our research goes beyond simply pursuing accurate modeling of SWE dynamics, by investigating the possibilities of using LSTM-based ML as an upper benchmark in the context of hypothesis testing (Gong et al., 2013; Nearing et al., 2020), that can be used to facilitate and guide improvements to *physically based* modeling of SWE dynamics. In Section 2, we introduce the LSTM-based and SN17 strategies for modeling snow, and discuss the data used for this study. Section 3 discusses the details of our experimental design. In Section 4, we present and discuss the results. In Section 5, we summarize our findings and discuss the outlook for future work.

2. Methods

2.1. Models

2.1.1. Long Short-Term Memory Network (LSTM)

An LSTM network is a type of recurrent neural network that includes memory cells that have the ability to store information over long time periods (Figure S1 in Supporting Information S1). These cells are subjected to three "gating" operations that effectively control the weight gradients and facilitate the learning of long-term dependencies (Hochreiter & Schmidhuber, 1997). Further, each memory cell functions in a manner analogous to a "state vector" in a traditional dynamical systems model, which makes the LSTM architecture an ideal candidate for developing models of dynamical systems (Kratzert et al., 2018); for a comprehensive hydrological interpretation of the LSTM architecture, please refer to Kratzert et al. (2018). In this study, we adopt the LSTM network architecture as used by Kratzert, Klotz, Shalev, et al. (2019) where the network architecture equations are summarized in the supplementary materials as Equations 1–7.

Table 1
Parameters for the SNOW17 Model Summarized by He et al. (2011a, 2011b) With Ranges Estimated From Anderson (1973)

Parameters	Explanation	Unit	Range
SCF	Snow fall correction factor	-	0.7–1.4
MFMAX	Maximum melt factor	mm per 6 hr per C^{o}	0.5–2.0
MFMIN	Minimum melt factor	mm per 6 hr per C^{o}	0.05–0.49
UADJ	The average wind function during rain-on-snow periods	mm per mbar per C^{o}	0.03–0.19
NMF	Maximum negative melt factor	mm per 6 hr per C^{o}	0.05–0.50
MBASE	Base temperature for non-rain melt factor	C^{o}	0.0–1.0
PXTEMP	Temperature that separates rain from snow	C^{o}	−2.0–2.0
PLWHC	Percent of liquid water capacity	-	0.02–0.3
DAYGM	Daily melt at snow-soil interface	mm d^{-1}	0.0–0.3
TIPM	Antecedent snow temperature index parameter	-	0.1–1.0

2.1.2. Snow Accumulation and Ablation Model (SNOW17)

The NWS is the US agency responsible for short-term and long-term streamflow predictions across the nation. The NWS RFC primarily uses the SNOW17 (hereafter SN17) model (Anderson, 2006) for generating operational forecasts of snow accumulation and melt in snow-dominated areas. SN17 is a spatially lumped process-based model that simulates snow accumulation and ablation. It requires three input data sets; air temperature and precipitation data are used as meteorological inputs, while information regarding elevation is used to compute atmospheric pressure. The model outputs include a rain-plus-snowmelt time series, as well as SWE (Figure S2 in Supporting Information S1). In this work, we apply the parsimonious point-scale assumption of full snow cover at the pixel-scale. Therefore, it is configured as a physical-oriented empirical temperature index model. We adopt the model structure and associated feasible parameter ranges (Table 1) presented by He et al. (2011a, 2011b).

2.2. Data

2.2.1. Meteorological Input Forcing

In this study, the LSTM and SN17 models were driven by meteorological forcing at the daily time scale. As inputs, we used daily values of precipitation, mean temperature, dewpoint temperature, and vapor pressure deficit from the Parameter-Elevation Regressions on Independent Slopes Model data set (PRISM; Daly et al., 1994). While PRISM data are more uncertain over complex terrain (Henn et al., 2018), it is arguably the best gridded climate data set available at this time, particularly for the western CONUS. We also used hourly, 0.125° near-surface downward longwave and shortwave radiation data from the near-real-time North American Land Data Assimilation Phase 2 data set (NLDAS-2; Xia et al., 2012). The hourly downward longwave and shortwave radiation data were first averaged to daily timescale and then resampled to 4-km resolution using nearest-neighbor interpolation onto the resulting grid coordinate with respect to PRISM.

2.2.2. Snow Water Equivalent (SWE) Target Variables

As the target variable for LSTM training and for SN17 calibration, and to evaluate the simulation results, we used the University of Arizona (UA) ground-based daily 4-km SWE data product (Broxton, Dawson, & Zeng, 2016; Broxton, Zeng, & Dawson, 2016; Zeng et al., 2018). This data set was developed by assimilating in-situ measurements of SWE and/or snow depth at thousands of sites (Broxton, Dawson, & Zeng, 2016; Broxton, Zeng, & Dawson, 2016; Dawson et al., 2018) using 4-km gridded PRISM precipitation and temperature data (Daly et al., 1994) over the CONUS. Accuracy and robustness of the UA snow product, and its use as a reference continental snowpack data set, have been assessed via four rigorous evaluation studies including point-to-point interpolation (Broxton, Dawson, & Zeng, 2016; Broxton, Zeng, & Dawson, 2016), pixel-to-pixel interpolation (Broxton, Dawson, & Zeng, 2016; Broxton, Zeng, & Dawson, 2016), and evaluation against independent snow cover extent data and airborne lidar measurements (Dawson et al., 2018). The UA SWE data product was found to align closely with the CONUS 1-km SWE product from the Snow Data Assimilation System (SNODAS; Barrett, 2003) and to show much better agreement with gamma SWE than the Special Sensor Microwave Imager

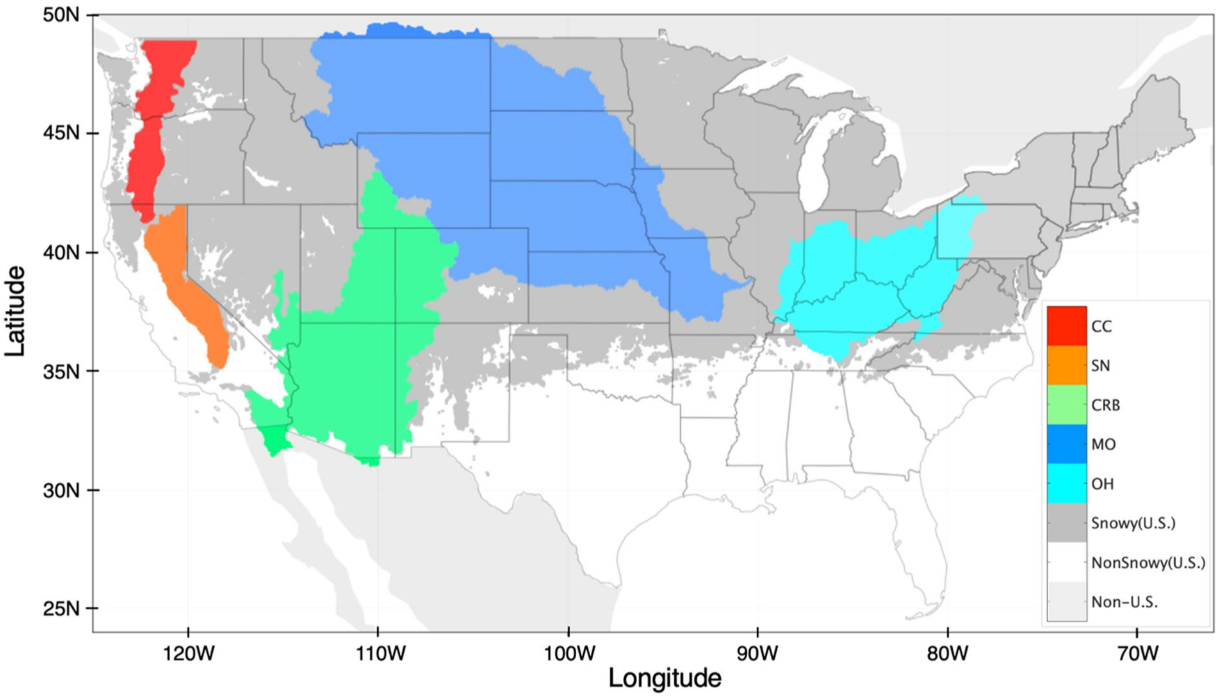

Figure 1. Geographic locations of five US regions include two Hydrologic Unit Code 2 (HUC2) basins (Ohio and Missouri) and three other regions in western CONUS (Colorado River Basin, Sierra Nevada, and Cascades). The "snowy" pixels are shaded gray and the non-snowy pixels are shaded white. OH = Ohio; MO = Missouri; CRB = Colorado River Basin; SN = Sierra Nevada; CC = Cascades. Note that, in this study, snowpack modeling is only performed within the CONUS, despite the fact that part of MO basin beyond the northern U.S. border and the part of CRB basin beyond the southern U.S. border are depicted in the figure.

and Sounder (SSMI/S) SWE and GlobSnow-2 SWE grid products for various land cover types and snow classes (Cho et al., 2020).

2.2.3. Static Features

To obtain gridded spatial maps of static land-surface characteristics, we used the open-source Geospatial Data Abstraction Library (GDAL)'s *gdal_translate* command-line tool to perform spatial reprojection of the Shuttle Radar Topography Mission (SRTM; Jarvis, 2008) digital elevation model elevation data to 4-km resolution; this is consistent with the PRISM latitude–longitude grid using average upscaling interpolation.

The "majority" operation was used to upscale the 1-km MODIS land cover climatology data set to the 4-km resolution grid (Broxton et al., 2014). We defined forested pixels to be those that included Evergreen Needleleaf, Evergreen Broadleaf, Deciduous Broadleaf, Mixed Forests, and Woody Savannas, whereas the remaining land cover types were classified as Non-Forested pixels.

3. Experimental Approach

3.1. Study Region

All the studies reported in this paper were conducted at 4-km pixel-scale over the CONUS, using a coordinate system and spatial coverage that is consistent with the PRISM meteorological forcing and UA SWE datasets. Roughly 0.31 million pixels were identified to be "snowy" from a total of approximately 0.46 million total pixels associated with the UA snow product data set. This categorization of snowy pixels was based on the snowpack climatology, where any pixel with median annual snowy season length less than 30 days, or median annual daily maximum SWE less than 10 mm, was classified as being "non-snowy," as shown in Figure 1 (Zeng et al., 2018).

Next, we selected five regions (shown in Figure 1) to explore the potential for using the LSTM architecture as a regional modeling tool, where the objective was to simulate snow accumulation and melt behavior over a large number of different pixels at the daily timescale. Three western CONUS regions—the Colorado River

Basin (hereafter CRB), Sierra Nevada (hereafter SN), and Cascades (hereafter CC)—were selected based on the important role that snowpack plays in contributing to their freshwater resources. The high elevation snowpack of the Rocky Mountains is known to contribute about 70% of the annual runoff of the Colorado River Basin (Christensen et al., 2004), while the Colorado River provides fresh water to over 40 million people in seven states and two countries (Deems et al., 2013). Also, in the SN and CC mountains (Simpkins, 2018), approximately 75% of freshwater originates from snow. To further ensure that the selected regions cover as wide a range of characteristics as possible in terms of geographic location, climatic regimes and local physiographic properties, we selected two additional USGS first level regions, namely Ohio (hereafter OH) and Missouri (hereafter MO), designated by a two-digit Hydrologic Unit Code (HUC).

The five selected regions cover a variety of topography and land cover regimes. The pixel aspect was derived from the SRTM digital elevation model (at its original resolution), and a consistent result was obtained by binning into eight directions; the two dominant aspects were determined to be north and south-facing slopes, together occupying around 30% of the total pixels over the five regions. For MO, the dominant aspect was determined to be north, and for the rest of the regions the dominant aspect was determined to be south. OH, MO and CC have mean elevations below 1,500 m (low elevation zone), while CRB and SN are between 1,500 m and 2,500 m with 18.2% and 16.3% pixels respectively having mean elevations above 2,500 m (high elevation zone; Mote, 2006). The OH, MO, and CRB have relatively lower percentages of forested pixels (about 36.6%, 3.81%, 9.69 respectively) whereas the SN and CC are recognized as being forest-dominated, with more than half of the pixels classified as forested (about 52% and 83%). These factors are known to exert strong controls on the energy balance during snowmelt (Garvelmann et al., 2015), and can be highly variable in space and time (Pohl et al., 2006).

3.2. Experimental Design

Data from the time period 1st October 1981 through 30th September 2000 were used for all model development runs—that is, SN17 calibration and LSTM training. For both the steps of calibration/training and testing of the models, we used data from the same time period, but from different spatially located pixels. In other words, our testing procedure evaluates the ability of each model to extrapolate in space, which is analogous to the problem of prediction in ungauged basins (Hrachowitz et al., 2013; Sivapalan et al., 2003). Note that it is computationally challenging (nearly impossible) to train either model (SN17 or LSTM) using data from the entire set of ~0.31 million snowy pixels; nor does it seem necessary. Instead, we use a process of sampling to select different, but representative, subsets of pixels to be used for training and for testing, as described in the following sections. As a precedent for this, Huo et al. (2019) have shown (in the context of sensitivity analysis) that the performance of a computationally intensive spatially distributed model can be reliably assessed by using only a sample of ~5% of the total number of pixels available over the CONUS (note that Huo et al., 2019 do not target SWE, but instead focus on gross primary production and latent heat flux, so their results might not necessarily be transferable).

To investigate the informational value of using different variables as input data, we developed 4 different LSTM models. To keep track of the different models, we adopt the following four-part naming convention, where the first part refers to the model type (LSTM or SN17), the second part refers to the Pixel Set used (A or B; these will be introduced later), the third part refers to the model domain (CONUS, Region or PX), and the fourth part refers to the variables used as input data (PT, PTE, 6M, and 6ME). In regard to the latter, CONUS refers to constructing a single model (learning a single set of parameters) that uses the input/output data from the entire country. Region refers to building separate models according to the number of selected regions. PX refers to achieving a separate "best-fit" model at each pixel. The abbreviation PT refers to precipitation and temperature, PTE refers to precipitation and temperature plus elevation, 6M refers to a set of six meteorological variables (precipitation, temperature, dew point temperature, vapor pressure deficit, longwave radiation, and shortwave radiation), and 6ME refers to the set of 6 meteorological variables plus elevation. For more detailed descriptions regarding the four-part naming convention for the LSTM networks and SN17 models used in this study, we refer the reader to Table S1 in Supporting Information S1.

The study reported here was conducted in several stages. In the first experiment (Section 3.2.1), we trained both model architectures (LSTM and SN17) to represent snowmelt dynamics at 15,000 pixels (*Pixel Set A*) selected randomly across the CONUS. This preliminary experiment had two purposes: (a) To determine whether the LSTM architecture is able to learn a better mapping relationship from inputs to outputs than the SN17 model, and

(b) To examine whether the LSTM network architecture is able to exploit the information provided by meteorological variables other than those used by SN17, to achieve better model performance.

Then, in the second experiment (Section 3.2.2), we evaluated both model architectures on a different set of 15,000 pixels (*Pixel Set B*; none of which were used in the first experiment) but selected in such a manner so as to provide equal representation to each of the five study regions mentioned above—OH, MO, CRB, SN, and the CC. The goals of this experiment were: (a) To assess whether the model performance obtained in the first experiment remains consistent when applied to another different data set, and (b) To examine the possibility of regional differences in performance.

In the third experiment (Section 3.2.3), we examined the transferability of LSTM-based models across regions. The goal was to investigate the extent to which different spatial regions share a common model structural representation.

3.2.1. Experiment 1: CONUS-Wide Modeling of Snow Accumulation and Melt

The purpose of the first experiment was to investigate the potential of using the LSTM machine-learning architecture as an alternative to the SN17 model structure for pixel-based CONUS-wide modeling of snow accumulation and melt. To this end, one single LSTM network was trained using input-output data from the entire country. Since training the network using data from more than 0.31 million pixels would be computationally prohibitive, we randomly selected 15,000 pixels from "snowy" areas across the CONUS (Figure 3). The goal was to obtain a representative subset of ~5% of the total number of possible snowy pixels. Then, to train the LSTM network, we constructed 15 bootstrap sample sets, each consisting of 1,000 different pixels randomly selected from the total set of 15,000 snowy pixels. We collectively refer to these 15 bootstrapped sets as *Pixel Set A*.

Accordingly, the four LSTM models developed for Experiment 1 were all CONUS-wide LSTMs trained on all pixels from *Pixel Set A*. For instance, *LSTM-A-CONUS-PT* is the model that was trained using only precipitation and mean temperature as forcing inputs; no information about local static pixel attributes (such as elevation, etc.) was used for development of this model. Similarly, the other three models are *LSTM-A-CONUS-PTE*, *LSTM-A-CONUS-6M*, and *LSTM-A-CONUS-6ME.*

As benchmarks for comparison, we developed two SN17 models including *SN17-A-CONUS* and *SN17-A-PX.* Note that SN17 currently uses only precipitation, temperature and elevation as input data. As such, the *SN17-A-CONUS* model can be thought of as representing a "lower-benchmark" on SN17 performance at each pixel, since this model treats all pixels as having identical functional characteristics, and simply applies the same input-state-output transformation algorithm to every pixel regardless of its location or local static characteristics. In contrast, the *SN17-A-PX* model can be thought of as representing an "upper benchmark" on SN17 performance at each of the calibrated pixels, since the model was tuned specifically to optimize performance at those pixels.

3.2.2. Experiment 2: Regional Modeling of Snow Accumulation and Melt

For this second experiment, we developed another set of 15,000 pixels, hereafter referred to as *Pixel Set B*, by randomly selecting 3,000 pixels from each of the five study regions (OH, MO, CRB, SN, and CC). Note that these five regions represent 13.11%, 3.53%, 13.79%, 67.93%, and 44.14% respectively of the total number of snowy pixels across the CONUS. Further, each region includes a different percentage of forested and non-forested areas. As a result, *Pixel Set B* has relatively dense representation of forested regions and is most sparsely representative of the MO river basin, which is the largest of the five regions.

Using *Pixel Set B*, we again trained the LSTM architecture at the CONUS level (a single model for the entire country), after which we trained separate LSTM models for each region (five separate models, one for each region). The procedure used was as follows. For each region, we partitioned the corresponding 3,000 pixels into sets of 1,000 each for training, validation, and testing. For the CONUS-wide model(s) the 1,000 "training" pixels from each of the five regions were grouped together to obtain 5,000 pixels to be used for network training (similarly for validation and testing). For each of the regional models, only the corresponding regional pixels were used for network training, validation, and testing.

To initialize each CONUS-wide LSTM model, we initialized the weights and biases using the corresponding results obtained at the Experiment 1; in other words, the network architectures were considered to have been "pre-trained" using the information provided by *Pixel Set A*. By doing so, we took advantage of the results of

Experiment 1 to minimize training costs, while achieving a consistent set of weights and biases for the CONUS-wide model that could be used when initializing the training of the separate regional models. In this way, we took advantage of the benefits of "transfer learning" (Kratzert et al., 2018). Accordingly, we developed four different LSTM models at the CONUS level, trained on *Pixel Set B* including *LSTM-B-CONUS-PT*, *LSTM-B-CONUS-PTE*, *LSTM-B-CONUS-6M*, *and LSTM-B-CONUS-6ME*.

Similarly, for each Regional LSTM model, we initialized the weights and biases using the corresponding results obtained from the CONUS-wide models trained on *Pixel Set B*; in other words, the regional network architectures were considered to have been "pre-trained" using the information provided by the CONUS-wide model trained on *Pixel Set B*. This approach took advantage of the results of CONUS-wide modeling to minimize training costs. Accordingly, we developed four different LSTM models for each Region including *LSTM-B-Region-PT, LSTM-B-Region-PTE, LSTM-B-Region-6M*, and *LSTM-B-Region-6ME.*

As benchmarks for comparison, we developed additional SN17 models including *SN17-B-CONUS*, *SN17-B-Region*, and *SN17-B-PX*. For model evaluation/testing, we focused on how well the models perform on the 5,000 testing pixels selected from *Pixel Set B* (1,000 pixels from each of the five regions). First, we evaluated the LSTM-based models against the SN17 benchmarks when using only PTE (precipitation, mean temperature, and elevation) as inputs, these being the same inputs used by SN17. The goal was to assess the capability of the LSTM network architecture to learn an appropriate representation of snow accumulation and melt over different training phases given the same input information that is available to the SN17 model. Then, we assessed which combination of inputs (PT, PTE, 6M, or 6ME) results in the best LSTM-based CONUS-wide predictions. Finally, we examined whether regional training results in better model performance than using the CONUS-wide model(s). Note that in all CONUS-wide cases, the LSTM-based models were fine-tuned on *Pixel Set B* after initializing using weights and biases trained on *Pixel Set A*.

3.2.3. Experiment 3: Exploring the Benefits of Transfer Learning

For the third experiment, we investigated the extent to which different spatial regions can share a common model structure with different parameter values through transfer learning (TL) across regions. Each LSTM-B-Region model trained in Experiment 2 was applied to each of the other four regions, resulting in 20 TL models for each input combination (PT, PTE, 6M, or 6ME). For instance, for each of the regions MO, CRB, SN and CC, the *LSTM-B-TL from OH* results are obtained by applying the four different TL-LSTM networks (trained with different input combinations on the OH Region) to those target regions. Similarly, we have the *LSTM-B-TL from Missouri*, *LSTM-B-TL from CRB*, *LSTM-B-TL from SN* and *LSTM-B-TL from Cascades* models.

The goal of this experiment was to test the extent to which a regional LSTM model structure hypothesis, imposed in the form of different kinds of regularization strategies at the input, can be transferred (extrapolated) to other locations. We benchmarked these TL-LSTM networks against the three models listed in Experiment 2 (*LSTM-B-Region, SN17-B-Region*, *and SN17-B-PX*) to examine how well the information about system structure extracted from one region can be transferred to another. To ensure a clean evaluation, the results were only assessed over the 5,000 testing pixels.

3.3. Objective Function

The objective function used for training the pixel-wise SN17 model was the Nash-Sutcliffe Efficiency (NSE; Nash & Sutcliffe, 1970) shown in Equation 1:

$$\mathrm{NSE} = 1 - \frac{\sum_{t=1}^{T}\left(\widehat{y_n^t} - y_n^t\right)^2}{\sum_{t=1}^{T}\left(y_n^t - \overline{y_n}\right)^2} \tag{1}$$

where $\overline{y_n}$ is the mean of observed SWE for each day, $\widehat{y_n^t}$ and y_n^t are the modeled and observed SWE at the training period time t $(1 \le t \le T)$ for a single pixel.

The objective function used for CONUS-wide and regional model training was $\mathrm{NSE}_{\mathrm{avg}}$ obtained by averaging the NSE values computed at each pixel that supplies training data (Kratzert, Klotz, Shalev, et al., 2019) shown as Equation 2:

$$\mathrm{NSE}_{\mathrm{avg}} = \frac{1}{P}\sum_{p=1}^{P}\sum_{n=1}^{N}\frac{\left(\widehat{y}_n - y_n\right)^2}{\left(s(p)+\epsilon\right)^2} \tag{2}$$

where P is the number of pixels, N is the number of days per pixel, $\widehat{y}_n$ is the prediction of pixel $n(1 \le n \le N)$, y_n is the observation, and $s(p)$ is the standard deviation of the SWE in pixel $p(1 \le p \le P)$, calculated from the training period. The value of $\epsilon$was set to 0.1 to avoid the loss function exploding to infinity for pixels with very low SWE variance.

3.4. Hyperparameter and Training Details

We mostly followed Kratzert, Klotz, Shalev, et al. (2019) for setting the LSTM hyperparameters; 256 hidden states, 1 stacked LSTM layer, a batch size of 256, a dropout rate of 0.4 and a sequentially decreased learning rate per 10 epochs from 1.0×10^{-3} to 5.0×10^{-4} then to 1.0×10^{-4}. The LSTMs were run in sequence-to-value mode, so that to predict a single daily SWE value the meteorological forcing from 242 preceding days, as well as the forcing data of the target day, were used (making the input sequences 243 time-steps long). This input sequence length follows suggestions from the land model community, where the snowy season is typically assumed to last from October 1st to May 31st resulting in a total of 243 days (Niu & Yang, 2007; Swenson & Lawrence, 2012). The relatively large number of hidden states (256) is believed to help circumvent the situation where the predictive performance of the LSTM is sensitive to weight initialization when using a small number of hidden state units (Bengio, 2012). The ADAM optimization algorithm was used for training (Kingma & Ba, 2014). Also, a single fixed random seed (2925) was applied to train all the LSTM networks. Our results indicated robust performance over three independent testing pixel sets (see Section 4.1.3), and therefore no further hyperparameter tuning was performed.

Note that in Experiment 1 we trained the LSTM network for a total of 15 epochs, where each epoch used data from a different bootstrapped set of 1,000 pixels taken from *Pixel Set A*. Here, an epoch refers to the LSTM training procedure wherein each temporal data sample for the entire set of 1,000 pixels is used once to update the values of the network parameters. The results of the 15th epoch were then used to initialize the training for Experiment 2, in which the LSTM network was trained for a total of 30 epochs using data from 5,000 training pixels selected from *Pixel Set B*. The results for the epoch having the highest average Kling-Gupta efficiency (KGE; Gupta et al., 2009) value over 5,000 validation pixels were then used to initialize the next stage of training for the five regional networks with 30 epochs each. Model performance was then assessed for an independent set of 5,000 testing pixels.

To calibrate the SN17 models, we used the *Shuffled Complex Evolution* (SCE) global optimization algorithm (Duan et al., 1992). Ten parameters were optimized, with the parameter range and model structure following He et al. (2011a). A standard batch calibration procedure was employed in which all training pixels were processed simultaneously at each iteration, in contrast to LSTM training where we randomly sampled pixels to make up each training batch to achieve faster convergence (LeCun et al., 2012).

3.5. Evaluation Metrics for Assessing Model Performance

To assess the consistency, reliability, accuracy, and precision of the models, we used several metrics, including NSE (Nash & Sutcliffe, 1970, Equation 2), the three components of KGE (Gupta et al., 2009) from Equations 3 to 6, and the scaled KGE (hereafter KGEss; Khatami et al., 2020, Equation 7):

$$\alpha^{\mathrm{KGE}} = \frac{\sigma_s}{\sigma_o} \tag{3}$$

$$\beta^{\mathrm{KGE}} = \frac{\mu_s}{\mu_o} \tag{4}$$

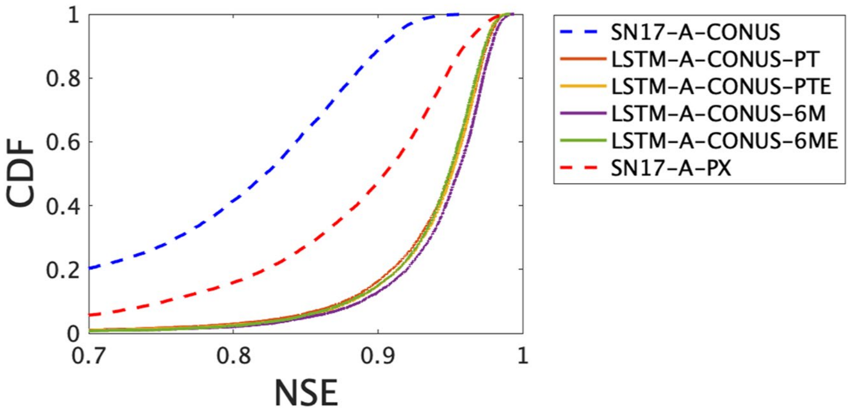

Figure 2. Aggregate Nash-Sutcliffe Efficiency performance for the Long Short-Term Memory networks (solid lines) and the benchmark SN 17 models (dashed lines) when applied to the 15,000 pixels from *Pixel Set A*.

$$\gamma^{\text{KGE}} = \gamma = \frac{Cov_{so}}{\sigma_s \sigma_o} \tag{5}$$

$$\text{KGE} = 1 - \sqrt{\left((\gamma^{\text{KGE}} - 1)^2 + (\beta^{\text{KGE}} - 1)^2 + (\alpha^{\text{KGE}} - 1)^2\right)} \tag{6}$$

$$\text{KGE}_{ss} = 1 - \frac{(1 - \text{KGE})}{\sqrt{2}} \tag{7}$$

where σ_s and σ_o are the standard deviation, and μ_s and μ_o are the mean of the simulated and observed SWE time series respectively, γ is the Pearson correlation coefficient and Cov_{so} is the covariance between the simulated and observed values.

4. Results and Discussion

4.1. Experiment 1: CONUS-Wide Modeling of Snow Accumulation and Melt

Figures 2 and 3 present a statistical assessment of the potential for using the LSTM architecture to model CONUS-wide snow accumulation and melt. The results show CDFs of testing-pixel performance over 15,000 pixels from *Pixel Set A* for the four CONUS-wide LSTM models (*LSTM-A-CONUS-PT, -PTE, -6M*, *and -6ME)* that use different input data sets, the lower-benchmark *SN17-A-CONUS* model, and the upper-benchmark *SN17-A-PX* model used as the bases for comparison. Note that each of these six models uses a single architecture to model snow accumulation and melt dynamics across the entire CONUS. Recall that the lower-benchmark *SN17-A-CONUS* model and the four LSTM models each use a single set of CONUS-wide parameters, while the *SN17-A-PX* model is individually trained to each pixel.

4.1.1. Comparative Overall Performance of the LSTM and SN17 Model Architectures

The NSE aggregate performance results (Figure 2) show clearly that the LSTM architecture provides better modeling of the general dynamics of snow accumulation and melt than the SN17 model architecture. All of the CDFs are shifted much further to the right, closer to the ideal value of 1.0. With a single network architecture and set of parameters applied to the entire CONUS, each of the four LSTM models (solid lines) achieves significantly better distributions of testing-pixel NSE scores than the lower- and upper-benchmark SN17 models (blue and red dashed lines respectively). In particular, the LSTM network architecture, with CONUS-wide sets of parameters (red, orange, purple, and green solid lines), provides better performance than the *SN17-A-PX* model for which parameters were optimized locally at each pixel. The unavoidable conclusion is that the SN17 model architecture does not adequately capture the structural nature of the input-state-output transformations that express the dynamics of snow accumulation and melt.

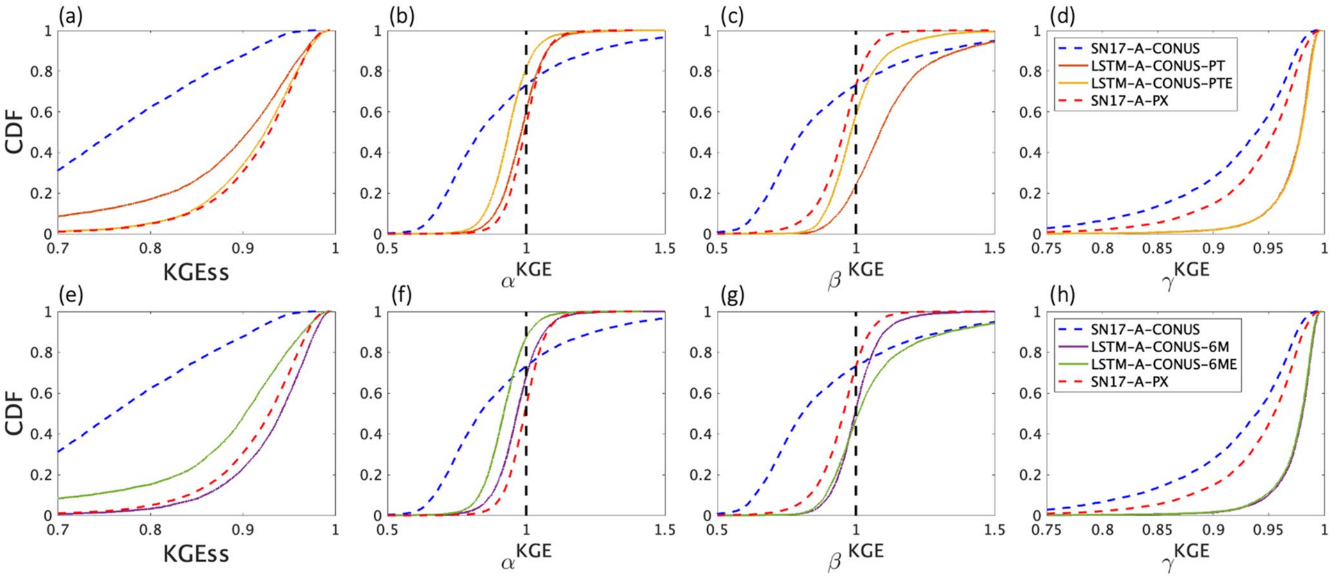

Figure 3. Aggregate performance, in terms of Kling-Gupta efficiency (KGE)ss and the three KGE components, for the Long Short-Term Memory networks (solid lines) and the benchmark SN 17 models (dashed lines) when applied to the 15,000 pixels from *Pixel Set A*.

4.1.2. Ability of the LSTM and SN17 Architectures to Exploit Information in the Input Data

Although the NSE metric indicates better aggregate performance of the LSTM architecture, it does not provide much insight into the reasons why. Note also that the use of different time periods to compute the aggregated NSE performance criterion can be informative (Schaefli & Gupta, 2007; Schaefli et al., 2005). Here, we use the $KGEss$ performance metric and its components to provide better discrimination between the models. The top row (a–d) of subplots in Figure 3 compares the results when both model types are provided with similar input data (precipitation, temperature, and elevation).

First, both the *LSTM-A-CONUS-PT* and *LSTM-A-CONUS-PTE* networks (red and orange solid lines respectively) achieve significantly better $KGEss$ performance than the *SN17-A-CONUS* model (blue dashed line). Further, the γKGE component shows that a major reason for this is that the LSTM is better able to simulate the shape and timing of snowmelt. So, even without any information regarding "local" properties of the landscape, the *LSTM-A-CONUS-PT* (which was not provided local elevation information) model is able to learn an input-state-output mapping that is better than that encoded by the SN17 model (which was provided with elevation information). In other words, the LSTM architecture is able to make better use of the information about snow dynamics provided by the input (precipitation and temperature) data.

Second, the *LSTM-A-CONUS-PTE* network *with* elevation information (orange solid line) is clearly better than the *LSTM-A-CONUS-PT* network *without* elevation information (red solid line). In particular, the use of elevation information results in a much better mass balance, as indicated by the β^{KGE} curve being closer to the ideal value of 1. This indicates, as might be expected, that there is considerable predictive value provided by the "local" information about elevation.

Third, the *LSTM-A-CONUS-PTE* network, with a single set of CONUS-wide parameters, achieves almost identical $KGEss$ performance to that of the *SN17-A-PX* model that was calibrated individually to each pixel (note that both these models use the same physical input information). This indicates that the LSTM architecture is able to successfully learn a set of parameters that enables it to be confidently applied to pixels that were not used for network training.

The subplots in Figure 3 compare the results when the LSTM network architecture is provided with different types of input data (red, orange, purple, and green solid lines). As indicated above, there is significant improvement when going from *PT* where only precipitation and temperature are provided (red solid line) to *PTE* where elevation information is also provided (orange solid line). However, even further improvement is achieved by the 6M network (purple solid line) that is provided with additional meteorological variables. Note that only the first two of these inputs (precipitation and temperature) are used by the SN17 model. So, providing the network with additional meteorological information (here dew point temperature, vapor pressure deficit, longwave radiation and shortwave radiation) is clearly beneficial. However, the 6ME model, which is provided with the six meteorological variables plus elevation, shows a clear decline in overall $KGEss$ performance. This seems to suggest that the information provided by elevation may be redundant when the meteorological information is provided (i.e., the meteorological variables already encode the useful information that would otherwise be provided by elevation).

Overall, the KGE metric and its components show that the *LSTM-A-CONUS-6M* model provides a much better representation of range of variability and water balance (curves are shifted closer to the center, where the ideal value is 1; Figures 3f and 3g), achieving a median $KGEss$ performance of 0.94 compared with the *SN17-A-PX* model (median $KGEss$ = 0.93). In general, all of the models tend to underestimate the variability of snowmelt ($\alpha^{KGE} < 1$; Figures 3b and 3g). Both the *LSTM-A-CONUS-PTE* and *LSTM-A-CONUS-6M* models provide better representations of snow mass balance (β^{KGE} closer to 1; Figures 3c and 3g) than the other 2 LSTMs. Use of only precipitation and mean temperature (*LSTM-A-CONUS-PT*) results in a tendency to positive bias, likely because it does not have access to humidity relevant information (vapor pressure deficit, dew point temperature) and is therefore unable to learn an accurate rain-snow partitioning threshold within the gating operation (Wang et al., 2019). Meanwhile, the *SN17-A-PX* model tends to underestimate mass balance suggesting that one should perhaps consider using other objective functions for pixel-wise training than NSE (or KGE, not shown, which results in similar underestimation bias).

In summary, the ability to exploit the information provided by a wider suite of meteorological variables enables the CONUS-wide implementation of the LSTM architecture to achieve a better representation of the dynamics of snow accumulation and melt, as assessed in terms of the ability to match the target SWE variable. However, even when provided with the same physical inputs as SN17, the LSTM architecture provides better results; such an implementation might be unavoidable when only Snow Telemetry (SNOTEL) information is available.

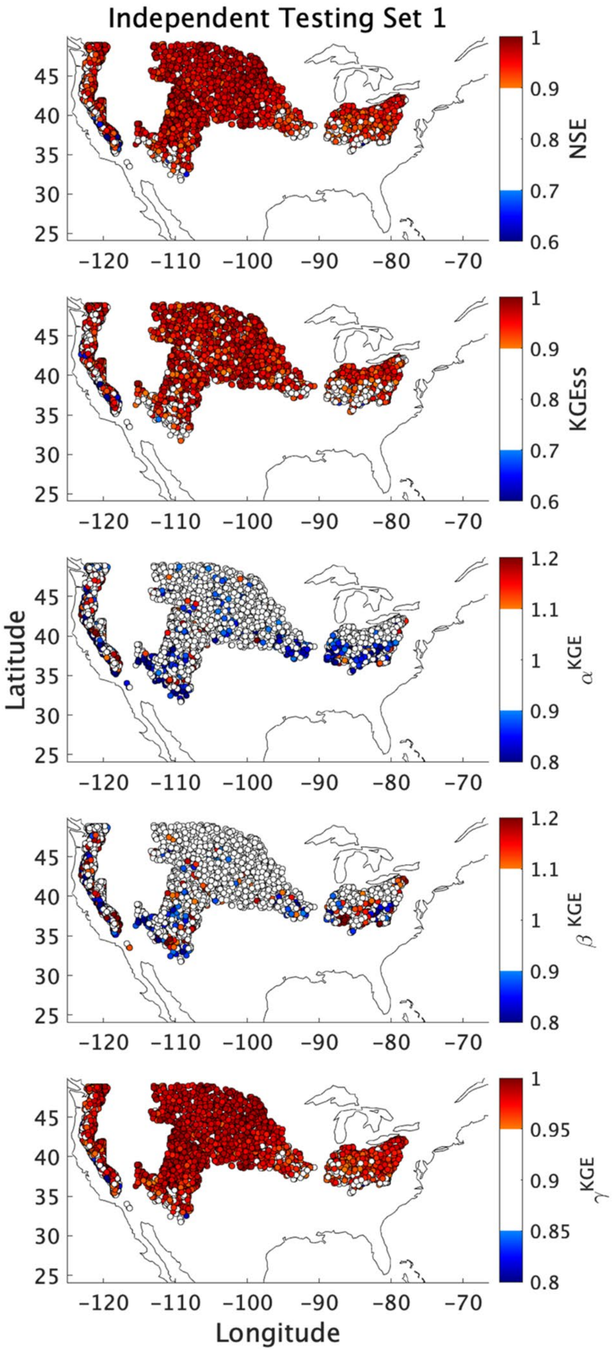

Figure 4. Spatial map indicating skill of the *LSTM-A-CONUS-6M* model (trained on *Pixel Set A*) when tested on an independent testing pixel set from *Pixel Set B*.

4.1.3. Evaluation on *Pixel Set B*

We next evaluated the *LSTM-A-CONUS-6M* model trained using *Pixel-Set-A* on a different set of 15,000 pixels from *Pixel Set B*. Figure 4 summarizes the spatial distributions of the five performance metrics separately for first independent sets (see other two sets in Figure S3 in Supporting Information S1) of 5,000 pixels from *Pixel Set B*. Table 2 shows that, overall, the model continues to provide good predictive performance on all three independent datasets, with only 31.69% and 30.22% of the pixels having more than ± 10% bias in the values of α^{KGE} and β^{KGE} respectively. Further detailed evaluation performed on each of the five regions (see Experiment 2) reinforced these findings (see Figure S1 in Supporting Information S1). Overall, these results indicate that the CONUS-wide *LSTM-A-CONUS-6M* model achieves a high degree of robustness with regard to predictive performance. As such, no further tuning of the LSTM model hyperparameters was performed in the later parts of this study.

The CDF plots related to Figure 4 are presented in Figure S4 in Supporting Information S1. Generally, MO and CRB have an overall better performance in terms of KGEss and NSE whereas the two forested areas, SN and CC, perform relatively worse, with OH being in between. A similar conclusion is found by examining the three KGE components, except that the OH shows a relatively large negative bias for the standard deviation error

(α^{KGE}). Table S2 in Supporting Information S1 summarizes the Pearson correlation coefficient (γ) for the pair of SWE hydrographs within the study period. MO and CRB have the two lowest average γ values for all the pairs (0.47 and 0.47 respectively) which suggests that the *LSTM-A-CONUS-6M* model trained over the entire country provides better performance when the region covers a more diversified climatic regime. Moreover, the model has worse average performance at forested (as opposed to non-forested) pixels, where the average KGEss skill difference between the two area equals 0.015, 0.079 and 0.023 over OH, SN, and CC respectively. Further, the KGEss skill is only 0.55 and 0.75 for Mixed Forest and Woody Savanna pixels in the SN region, suggesting the need to either construct separate local models, or to add relevant local attributes to the CONUS-wide models to improve overall LSTM model performance.

Table 2
Summary Statistics for the LSTM-A-CONUS-6M Model Evaluation Results Over Pixel Set B

Evaluation over *Pixel Set B*				
Model skill	Percentage of pixels for independent Test Set 1	Percentage of pixels for independent Test Set 2	Percentage of pixels for independent Test Set 3	Percentage of pixels for all test sets
$\mid 1 - \beta^{\mathrm{KGE}}\mid > 10\%$	31.92%	31.74%	31.40%	31.69%
$\mid 1 - \beta^{\mathrm{KGE}}\mid > 10\%$	30.24%	30.20%	30.22%	30.22%
$\gamma^{\mathrm{KGE}} > 0.95$	82.08%	82.42%	80.92%	81.81%
$\mathrm{KGE}_{\mathrm{ss}} > 0.95$	32.72%	30.84%	31.06%	31.54%
NSE > 0.95	41.18%	40.34%	41.84%	41.12%
$\gamma^{\mathrm{KGE}} < 0.85$	2.20%	1.96%	1.88%	2.01%
$\mathrm{KGE}_{\mathrm{ss}} < 0.70$	3.96%	3.90%	4.22%	4.03%
NSE < 0.70	3.72%	3.70%	3.70%	3.71%

4.2. Experiment 2: Regional Modeling of Snow Accumulation and Melt

4.2.1. Results of CONUS-Wide Fine Tuning

Experiment 2 was conducted in stages, where the first stage was another round of CONUS-wide network training. In the following discussion, we refer to the LSTM networks obtained by training on *Pixel Set A* as the "pre-trained" CONUS-wide networks. Initialized from the weights and bias parameters of these pre-trained networks we conducted a further stage of CONUS-wide network training using *Pixel Set B* (called the "fine-tuned" CONUS-wide networks). The results of this second round of network training are the *LSTM-B-CONUS-PT*, *-PTE*, *-6M*, and *-6ME* CONUS-wide models, as described in Section 3.2.2.

Performance comparison of the pre-trained (using *Pixel-Set-A)* and the fine-tuned (using *Pixel-Set-B*) CONUS-wide LSTM networks is shown in Figure 5. The results show performance on the 5,000 independent testing pixels from *Pixel Set B*. Note that while, performance was already quite good based on *Pixel-Set-A* training, the model skill, as measured by the median value of KGEss, improves by ∼0.08/∼0.06/∼0.08 for the *PT*/*PTE*/*6ME* models respectively, and by only ∼0.01 for the *6M* model. This reinforces the earlier finding that use of a full suite of meteorological variables results in an efficient basis for training the LSTM network. However, providing the network with additional information about elevation (*-6ME* model) does not result in further improvement.

This added value of fine-tuning is further demonstrated in Figure 6, which shows the change in model skill from *LSTM-A-CONUS-6M* to *LSTM-B-CONUS-6M.* The left column of subplots shows the geographical distribution of change in model skill (blue indicates improvement) while the right column shows the corresponding performance difference CDFs individually for each of the five regions. In the right column, the metric α_*^{KGE} is defined as $1 - |1 - \alpha^{KGE}|$ and the metric β_*^{KGE} is defined as $1 - |1 - \beta^{KGE}|$ so as to better illustrate the change in skill. Accordingly, positives values in the right column of subplots indicate improved performance of *LSTM-B-CONUS-6M* over *LSTM-A-CONUS-6M* with respect to the corresponding metric.

Accordingly, of the 5,000 testing pixels, 58.0% have improved NSE while 57.8% have improved KGEss (with 59.6%, 54.2% and 55.8% improvement for the α^{KGE}, β^{KGE}, and γ^{KGE} components respectively). More than half of the pixels show improvements for NSE and all three components of KGE for the regions other than OH. In OH, as many as 60% of the pixels show a decrease in KGEss skill (due to γ^{KGE}). This may be because *Pixel-Set-A* contains a larger number of non-forested snowy pixels (77%) over the CONUS than *Pixel-Set-B* (62%). Since the CONUS-wide model has to select network parameters that balance performance over both forested and non-forested areas, the result seems to be improved over forested areas (which are better represented *by Pixel-Set-B*) at the expense of non-forested areas.

4.2.2. Regional Training of the LSTM Models

Initialized from the weights and bias parameters of the fine-tuned CONUS-wide networks we next trained a separate LSTM network for each of the five regions, again using *Pixel-Set-B* (we refer to these as the "fine-tuned" regional networks). Overall, at the CONUS-level, the regional tuning results in the median CDF of KGEss

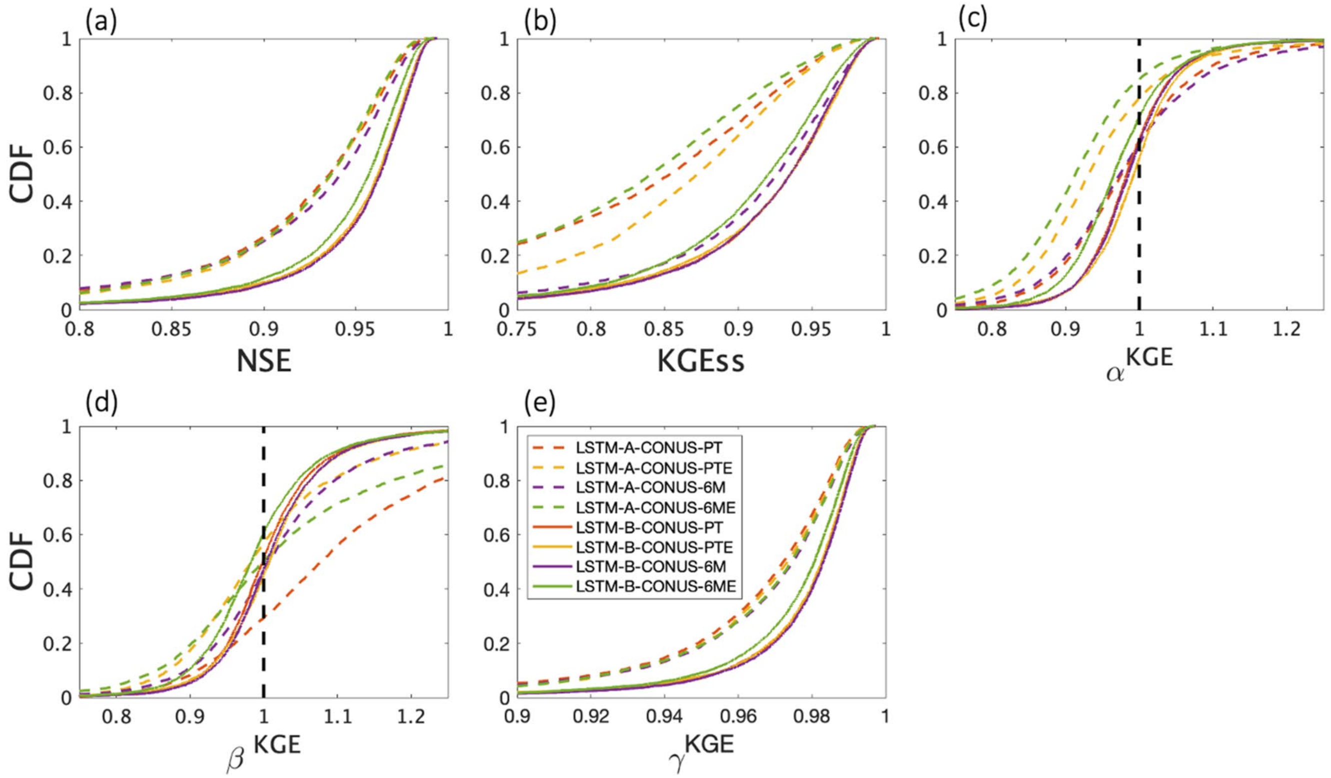

Figure 5. Aggregate performance of the trained CONUS-wide Long Short-Term Memory networks after fine-tuning using *Pixel Set B* compared to when pre-trained using *Pixel Set A*, where the evaluation is conducted over 5,000 independent testing pixels from *Pixel Set B*.

improving by a small amount—by 0.013, 0.012, 0.013, and 0.025 for the PT, PTE, 6M, and 6ME models respectively (see Figure S5 in Supporting Information S1). While the *6ME* model shows the largest improvement, its overall performance is still worse than for the other models (consistent with previous results).

In contrast with the CONUS-wide fine-tuning stage (Figure 6), regional fine-tuning results in even more improvement of model skill across the five regions (Figure 7). The range of improvement is from 55% (SN) to 65% (MO) for α^{KGE}, from 56% (SN) to 62% (CRB) for β^{KGE}, and from 77% (SN) to 88% (OH) for γ^{KGE}. Overall, 81% of the testing pixels show improved NSE skill, while 64% show improved KGEss (60%, 59% and 81% for α^{KGE}, β^{KGE}, and γ^{KGE} components respectively). Meanwhile 85% (68%) of the forested pixels show greater improved NSE (KGEss) skill than the CONUS-wide fine-tuning 54% (51%) over the SN region. A general conclusion is that allowing the LSTM network to account for regional differences helps improve predictive performance, especially over forested areas.

4.2.3. Comparison With SN17 Benchmarks

The results reported above indicate that the best performing model is the *LSTM-B-Region-6M* deep learning network architecture trained separately to each region. In this section, we evaluate the extent to which the LSTM architecture is able to “learn” a better input-output mapping than is encoded by the SN17 model, when both modeling strategies are provided with the *exact same* input information (precipitation, temperature, and elevation) over different phase of model development. Figure 8 summarizes the progression of performance of the LSTM architecture (evaluated over the 5,000 *Pixel-Set-B* testing pixels), starting with the pre-trained *LSTM-A-CONUS-PTE*, proceeding to the fine-tuned *LSTM-B-CONUS-PTE*, and finally to the five fine-tuned *LSTM-A-Region-PTE* models (here grouped together as one larger CONUS-wide model with regional differentiation). As benchmarks for comparison we show the *SN17-A-CONUS* model (black dashed line) and corresponding *SN17-B-CONUS* model (red dashed line), each of which uses a single set of parameters to represent the entire CONUS, the *SN17-B-Region* model (blue dashed line) that uses five different parameter sets (one set for each of the five regions), and an “upper-benchmark” *SN17-B-Pixel* model (black dotted line) that is individually calibrated to

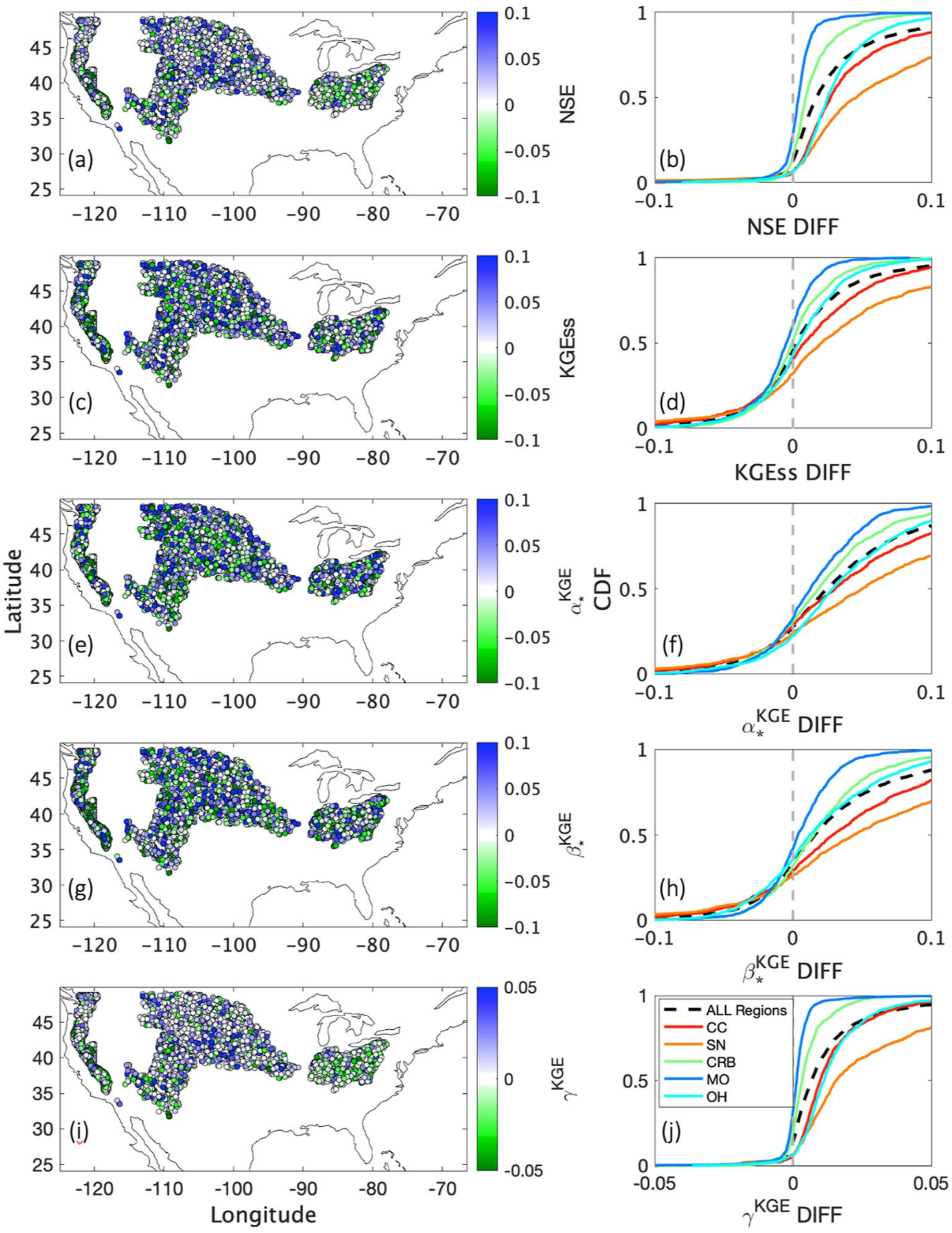

Figure 6. Difference in model skill between the CONUS-wide Long Short-Term Memory (LSTMs) trained on *Pixel Set B* and *Pixel Set A*, when using the six meteorological variables, evaluated over the 5,000 testing pixels from *Pixel Set B*. Note that $\alpha_*^{KGE} = 1 - |1 - \alpha^{KGE}|$, $\beta_*^{KGE} = 1 - |1 - \beta^{KGE}|$. Movement of the CDFs to the right (to more positive values) indicate that the *LSTM-B-CONUS* models have better performance than the corresponding *LSTM-A-CONUS* models. OH = Ohio; MO = Missouri; CRB = Colorado River Basin; SN = Sierra Nevada; CC = Cascades.

each of the 5,000 testing pixels (thereby reflecting the best possible performance achievable at those pixels by the SN17 model architecture given the available data).

First, we notice that, as might reasonably be expected, the SN17 and LSTM models get progressively better (in terms of all of the reported metrics) as we proceed from the *CONUS* to *Regional* versions. However, this progressive improvement is much more significant for the SN17 model (see KGEss, α^{KGE} and β^{KGE} metrics) than for the LSTM models. Further, the LSTM architecture has learned a far better representation of the snow-accumulation and melt input-output mapping than is expressed by the SN17 model architecture. In terms of the NSE metric,

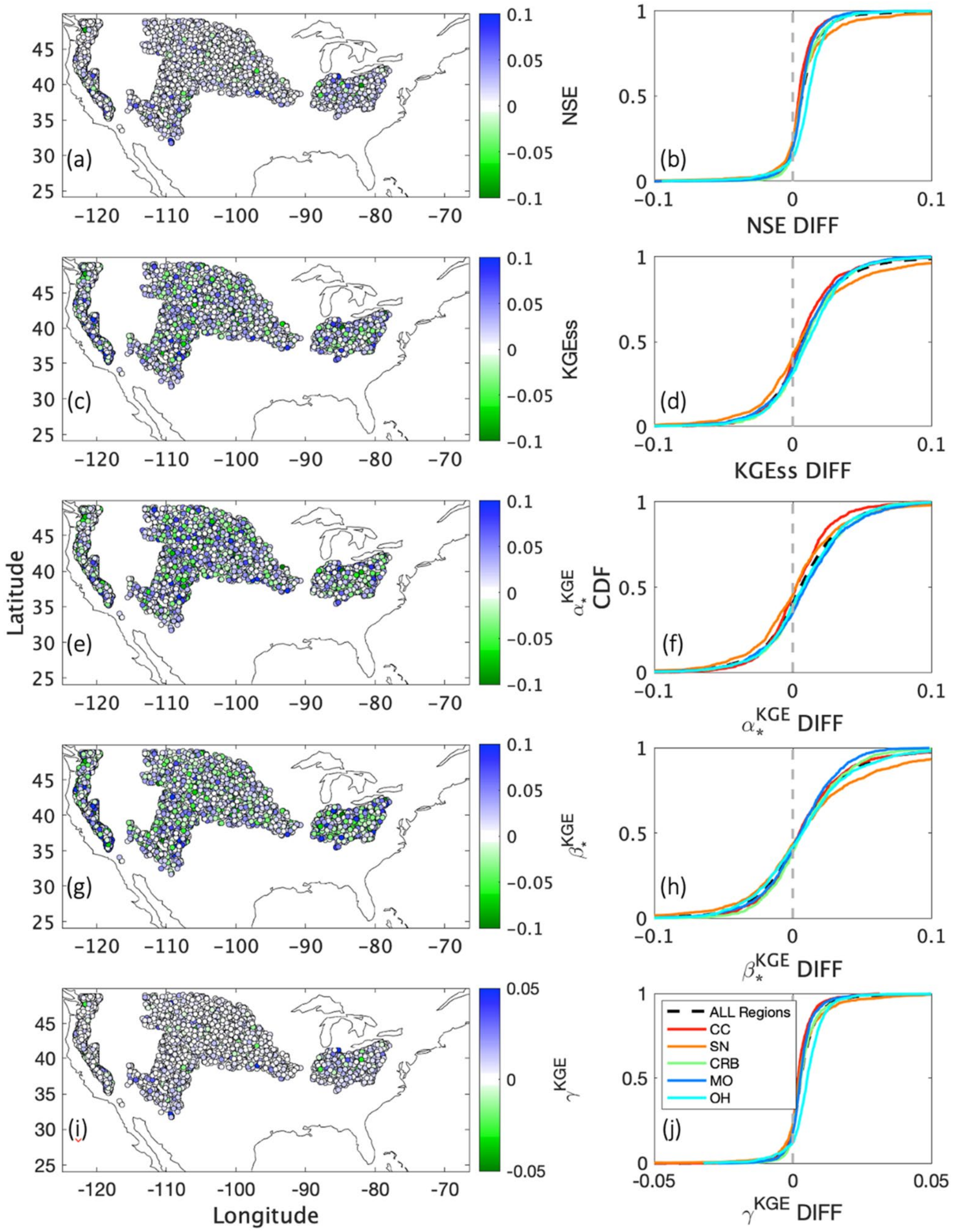

Figure 7. Difference in model skill between the regional Long Short-Term Memory (LSTM) and CONUS-wide LSTMs trained on *Pixel Set B*, when using the 6 meteorological variables, evaluated over the 5,000 testing pixels from *Pixel Set B*. Note that $\alpha_*^{KGE} = 1 - |1 - \alpha^{KGE}|$, $\beta_*^{KGE} = 1 - |1 - \beta^{KGE}|$. Movement of the CDFs to the right (to more positive values) indicates that the *LSTM-B-Region* models have better performance than the corresponding *LSTM-B-CONUS* models. OH = Ohio; MO = Missouri; CRB = Colorado River Basin; SN = Sierra Nevada; CC = Cascades.

all three LSTM models (*A-CONUS*, *B-CONUS*, and *B-Regional*) achieve median NSE values above 0.95, while the best comparable SN17 model (*SN17-B-Regional*) achieves a median NSE value of around 0.82. Note that the *SN17-B-Pixel* results, which represents a "best possible" SN17 model since the model was calibrated to the testing pixels is still worse (with a median NSE value of around 0.87) than all three of the LSTM models. In contrast, the KGEss metric indicates that the pre-trained *LSTM-A-CONUS-PTE* model is only slightly better than the *SN17-B-Region* model and worse than the *SN17-B-Pixel* benchmark. Meanwhile the *LSTM-B-CONUS-PTE* and *LSTM-A-Region-PTE* models have the best performance.

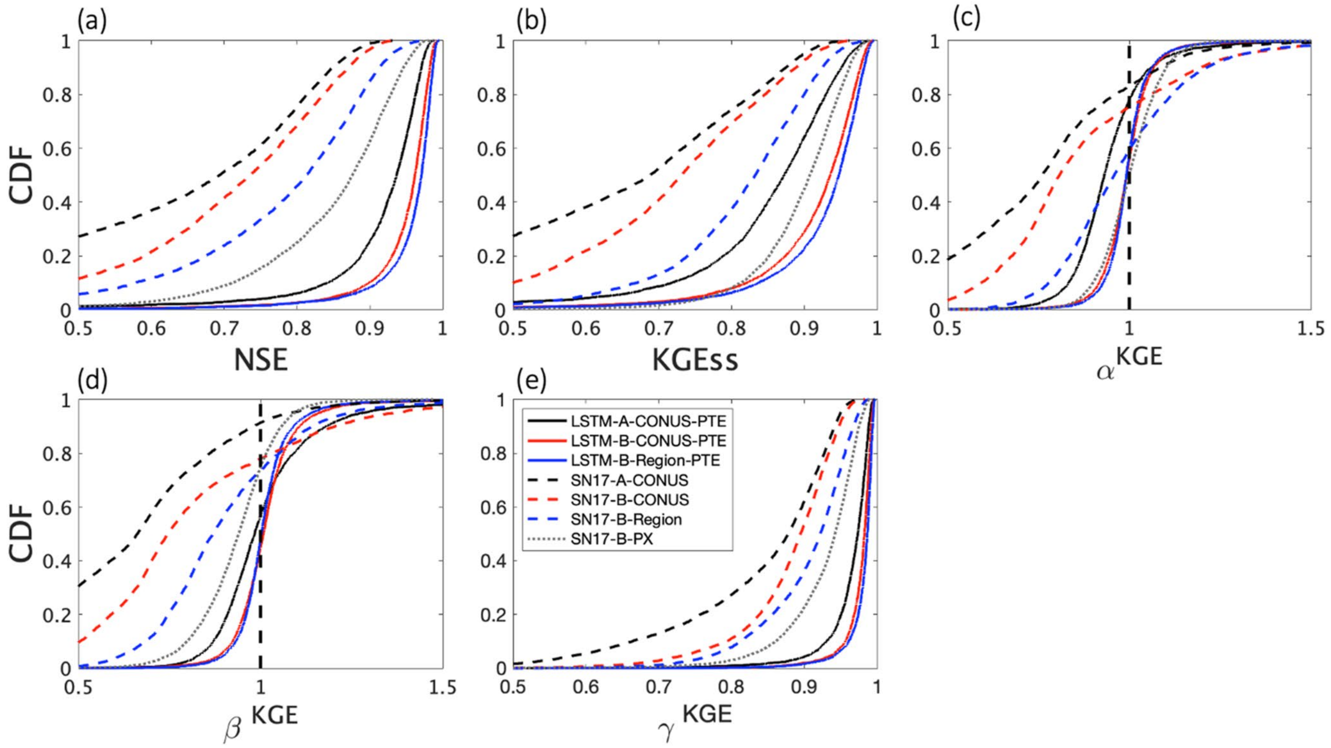

Figure 8. Aggregate performance of the Long Short-Term Memory models (solid lines) benchmarked against the SN17 models (dashed lines) when both are given the same input information (precipitation, temperature and elevation), evaluated over the 5,000 testing pixels from *Pixel Set B*.

Finally, it should be noted that although the improvement from *LSTM-B-CONUS-PTE* to *LSTM-A-Region-PTE* is both clear and consistent, it is not very large; this indicates that a trained CONUS-wide LSTM model (based on PTE data) is capable of providing almost as good performance as a regionally trained one. Further this CONUS-wide LSTM is better than the regionally trained SN17 model and is even better than the "optimized" SN17-B-Pixel model that was calibrated to achieve best possible performance at the "testing" pixels; this latter finding is consistent with the "prediction in ungauged basins" results reported by (Krazert, Klotz, Herrnegger, et al., 2019; Krazert, Klotz, Shalev, et al., 2019) in the context of rainfall-runoff modeling.

4.2.4. Some General Remarks

In general, the relatively good performance of the LSTM-based models should (perhaps) not be too surprising since it is likely that a much larger amount of information has been assimilated by the deep learning process than was available to the developers of the SN17 model architecture. What does seem remarkable is that the collective-regionally-differentiated ("fine-tuned" CONUS-wide) LSTM model is not very much better than the single CONUS-wide representation, suggesting that the latter *may* be capable of providing acceptably good predictions of SWE at locations that are not necessarily similar, in terms of local attributes, to the conditions experienced by the model during training; in other words, *the conditions determining the dynamics of snow accumulation and melt depend largely on meteorological drivers*, *whereas the local conditions may have only marginal impact*, at least at the scale of the individual pixels used for this study.

We illustrate this point by examining mass balance error (i.e., ($|1-\beta^{KGE}|$)) as summarized in Table 3. For example, the regionally differentiated performance of the CONUS-wide (*LSTM-B-CONUS*) model drops only slightly in MO and CRB, where the average error drops by merely 0.34% (from 4.85% to 4.51%) and 2.02% (from 6.99% to 4.97%). The single CONUS-wide representation clearly has higher transferability between non-forested regions even though they may include a wide range of climate regimes in terms of latitude coverage.

As mentioned earlier, accounting for regional differences via sampling helps improve predictive performance over forested areas. Here, we observe significant reduction in bias (by 8.11% and 4.30% respectively) in SN and

Table 3

Summary Mass Balance Error (|1 − β^{KGE}|) for the LSTM-A-CONUS-6M, LSTM-B-CONUS-6M, and LSTM-B-Region-6M Models Evaluation Results Over Pixel Set B

LSTM-A-CONUS-6M (total 15,000 testing pixels)					
Models/Regions	OH	MO	CRB	SN	CC
Avg. of $\lvert 1-\beta^{KGE}\rvert$	8.01%	4.85%	6.99%	17.35%	10.40%
Max. of $\lvert 1-\beta^{KGE}\rvert$	132.86%(4)	41.41%	58.57%	240.61%(8)	74.64%(1)
No of pixel $\lvert 1-\beta^{KGE}\rvert \geq 1.0$	1	-	-	49	-
No of pixel $\lvert 1-\beta^{KGE}\rvert \geq 2.0$	-	-	-	3	-
LSTM-B-CONUS-6M (total 5,000 testing pixels)					
	OH	MO	CRB	SN	CC
Avg. of $\lvert 1-\beta^{KGE}\rvert$	5.68%	4.52%	4.97%	9.24%	6.10%
Max. of $\lvert 1-\beta^{KGE}\rvert$	39.65%	38.84%	38.25%	99.58%(1)	91.06%(5)
LSTM-B-Region-6M (total 5,000 testing pixels)					
	OH	MO	CRB	SN	CC
Avg. of $\lvert 1-\beta^{KGE}\rvert$	4.85%	4.07%	4.07%	7.85%	5.20%
Max. of $\lvert 1-\beta^{KGE}\rvert$	32.68%	35.63%	31.86%	85.29%(1)	63.38%(5)

Note. The integer adhere to the large maximum mass balance error within the bracket refers to the International Geosphere-Biosphere Programme (IBGP) land classes as presented in Broxton et al. (2014). 1: Evergreen needleleaf; 2: Evergreen broadleaf; 4: Deciduous broadleaf; 5: Mixed forests; 8: Woody savannas.

CC. The maximum bias in OH and SN also decreases considerably (by 93.31% and 141.03% respectively). The improvement in SN likely comes from the large total number of local pixels in *Pixel Set B*. Overall, significant decreases of maximum mass balance error in SN (by 14.29%) and CC (by 27.68%) during the regional fine-tuning indicates the importance of exploiting the local meteorological information when predicting forested region snowpack. The remaining large value of the maximum bias (SN: 85.29%; CC: 63.38%) suggests that the model may need to be provided with local information at a finer spatial resolution, and that the current LSTM gating function may have difficulties learn the effects of vegetation interception when provided only the 6M meteorological context.

4.2.5. Time Series Illustrating Daily Snowpack Dynamics

Figure 9 shows that the regional LSTM is able to properly simulate the seasonal cycle dynamics of snow accumulation and melt quite well. The figures show, for 10%, 50%, and 90% percentiles of KGEss, time-series comparisons of simulated and observed SWE selected from *Pixel Set B* testing pixels for the *LSTM-B-Region-6M* model, *SN17-B-Region*, and *SN17-B-PX* model. Additional examples for the CONUS-wide LSTM and other versions of the SN17 model are presented in Figure S6 in Supporting Information S1 with the associated KGEss statistics and pixel information summarized in Table S3 in Supporting Information S1. Although a very large number of cases was investigated in this study, the results presented here can be considered to be representative and hence useful for understanding how informative the chosen metrics are about the characteristics of any given simulation.

Further, we broaden the list of statistical goodness-of-fit measures to include the normalized total SWE bias, as well as a timing measure related to the reproduction of annual day of peak SWE at the pixel scale. From Figure S7 in Supporting Information S1, we see that the LSTM clearly has smaller total SWE biases compared to the corresponding different SN17 models. Notice again that a CONUS-wide LSTM strategy (e.g., *LSTM-A-CONUS-6M*) would only perform better than the regionally fine-tuned model in terms of total normalized bias for the MO basin because the former covers a more diversified snowpack regime (Table S3 in Supporting Information S1).

On the other hand, the LSTM does not reproduce the occurrence of annual maximum SWE date better than the SN17 models (Figure S8 in Supporting Information S1) Note that SN17 is also not able to reproduce this important timing measure. Given that our current LSTM development strategy uses a fixed number of 243 past time-steps as the input sequence length, it would be useful to explore the benefits of varying the memory window size as a possible way of enhancing the ability of the LSTM to better capturing timing-relevant signatures of snow dynamics.

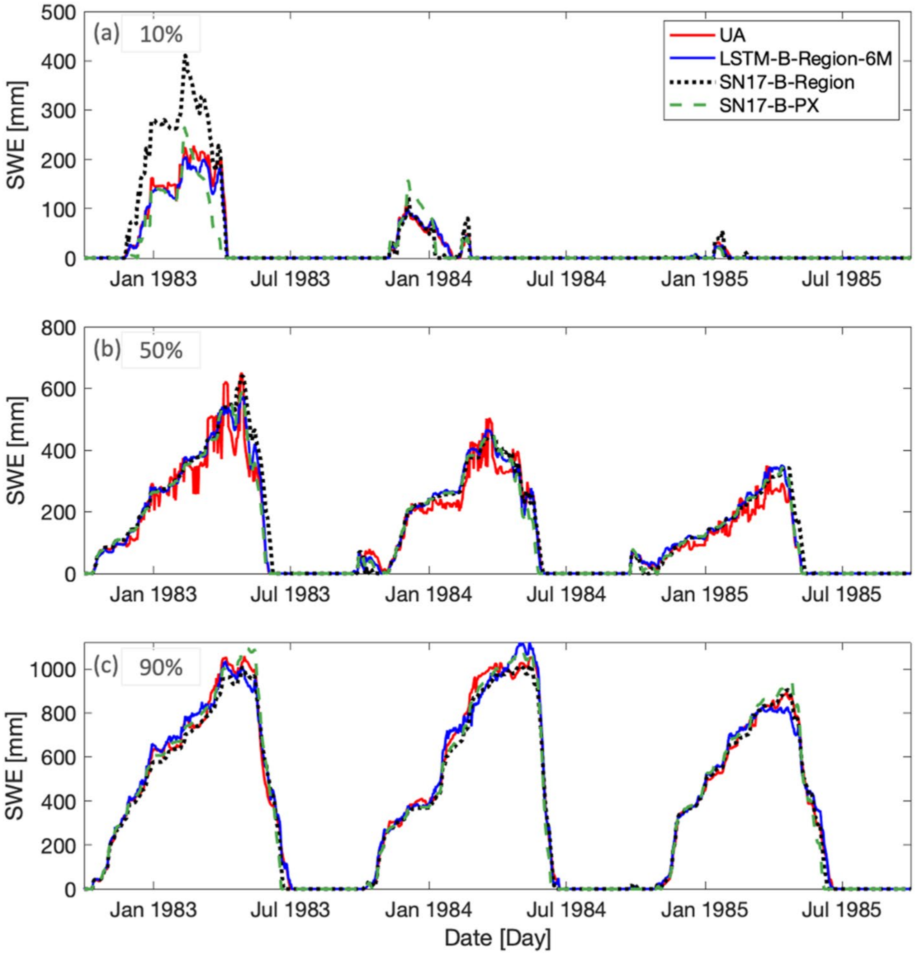

Figure 9. Comparisons between UA observations and SWE predicted by *LSTM-B-Region-6M*, *SN17-B-Region*, and *SN17-B-PX* models. The pixels are from 5,000 testing pixels in *Pixel Set B* and represent locations corresponding to the a) 10th, b) 50th, and c) 90th percentiles of KGEss performance of the model. The results are shown from WY1983 to WY1985.

4.3. Experiment 3: Exploring the Benefits of Transfer Learning

In Experiment 1, we demonstrated the ability of a CONUS-wide LSTM to make accurate and robust predictions at continental scale, across different pixel sets. In Experiment 2, we showed that a regionally trained LSTM also shows promising performance when tested on independent pixels within the same region. Here, we explore the potential for transfer learning (TL), in which we evaluate the extent to which an LSTM trained to one region can be used outside of the original regional for which it was developed. This is achieved by applying the regional LSTM network (*LSTM-B-Region*) trained from one region to the remaining four corresponding regions and evaluating performance on 1,000 testing pixels selected (within that region) from *Pixel Set B*. The evaluation results in total 20 TL evaluations for each type of LSTM (PT, PTE, 6M, 6ME) that uses different input information.

4.3.1. Evaluation Metric for Transfer Learning

To quantify the KGEss performance of each regional TL-LSTM network, we compute the area under the CDF curve integrated between 0 and 1 (i.e., positive values of KGEss). We then obtain the $\phi_{S\to T}$ TL metric by subtracting the integrated area from 1.0 as shown by Equation 8:

$$\phi_{S\to T} = 1 - \int_0^1 f_{KGEss(S\to T)} d\,(KGE_{ss}) \tag{8}$$

The symbol S refers to the source region where the TL network was developed, T refers to the target region that the network is applied to, and f_{KGEss} indicates the KGEss performance CDF of the TL network. Because KGEss performance is generally positive for all of the cases examined, we neglect area under the CDF curve corresponding to KGEss less than 0. Accordingly, the $\phi_{S\rightarrow T}$ metric is bounded between 0 and 1 with larger values indicating better TL performance. Finally, the degree of transferability of each regional LSTM is evaluated as the ratio written as Equation 9:

$$R_{S\rightarrow T} = \frac{\phi_{S\rightarrow T}}{\phi_T} \tag{9}$$

where ϕ_T refers to the metric (Equation 8) computed for the model when trained specifically to the target region (i.e., not transferred). Accordingly, we compare the regional TL-LSTM networks against three benchmarks, including the regional LSTM model trained to the target region ($\phi_{T_{LSTM}}$), the regional SN17 model trained to the target region ($\phi_{T_{SN17-Region}}$) and the SN17 pixel model trained to the target region ($\phi_{T_{SN17-PX}}$). Thus, values of $R_{S\rightarrow T}$ larger than 1.0 indicate that the TL model is able to outperform the benchmark model, while values less than one indicate poor ability to exploit transfer learning.

4.3.2. Evaluation of Regional TL Networks Against Three Benchmarks

We first compare the TL-LSTMs against the target region LSTMs (see Figure 10). We put less attention on the results of 6ME network (Figure S8 in Supporting Information S1) than the PT, PTE, and 6M network because it has the worst performance while requiring more input information. In general, this suggests that "elevation" information can be considered to be redundant when the set of six meteorological variables is made available. Overall, we see that the *LSTM-PT* networks, which require fewer input data provide better TL performance (Figure 10a) than the PTE (Figure 10b) and 6M networks (Figure 10c), as indicated by the majority of the $R_{S\rightarrow T}$ values being close to or larger than 1.0. This suggests that the LSTM-gating operations have been able to learn a better universal representation of the processes that control the rain-snow partitioning and snowmelt dynamics, by exploiting only the information provided by precipitation and mean temperature. This finding also suggests the existence of a tradeoff between model transferability and model complexity (in the sense of the number of input variables used for training the LSTM network; Lute & Luce, 2017). However, whether this finding is general requires further investigation and consideration of issues such as data quality and quantity (Schoups et al., 2008).

As shown by Figures 10b and 10c, the TL results deteriorate when elevation is included as an input for LSTM network training. In particular, the $R_{S\rightarrow T}$ metric for *LSTM-B-TL from CRB* decreases to 0.88 (6M) and 0.65 (PTE) when applied to Ohio, while conversely the *LSTM-B-TL from Ohio* decreases to 0.92 (6M) and 0.57 (PTE) when applied to CRB. So, the inclusion of elevation as training information tends to cause the LSTM to learn a regional representation that does not transfer well to other regions.

Similarly, the PTE and 6M versions of the *Cascades LSTM-B-TL* do not transfer well to the other four regions, and especially to the three non-forest regions (for which the $R_{S\rightarrow T}$ values are all less than 0.50). This makes sense, since the Cascades is a relatively unique region where the spatial coverage is relatively narrow and is located at higher latitudes, so that a representation is learned that is locally specific and therefore does not transfer well to geographical regions that are not similar. However, this result may also be due to the way that we have fine-tuned the LSTM network, because the 6ME results (Figure S9 in Supporting Information S1) also transfer poorly from *CRB* and *SN* to *OH* and *MO*. Because we have allowed all of the weights and biases to be tunable at each level of network training, the regional CC networks for PTE and 6M have likely forgotten some of the general learning achieved through CONUS-level training. It is possible that freezing some of the weight and biases during regional training may help to address the reasons for poor network transferability (Ma et al., 2021).

Next, we compare the TL-LSTMs to the target region SN17 benchmarks (SN17-Regional and SN17-PX). We see that about 80% (Figure 10d). And more than half (55%; Figure 10g) of the TL-LSTMs that used only precipitation and mean temperature as inputs outperform (indicated by the blue-green color) the corresponding target region SN17-Regional and SN17-PX models. When using all six meteorological inputs (6M) this success rate decreases to only 64% (Figure 10f) and 52% (Figure 10i) against the target region SN17-Regional and SN17-PX models respectively. Although the TL-LSTM using PTE use the exact same type of input information as SN17, its transferability shows a further decrease to 52% and 36%. So, although the LSTM may not be fully exploiting the information provided by the elevation data, LSTMs trained for other regions are still able (to a certain extent)

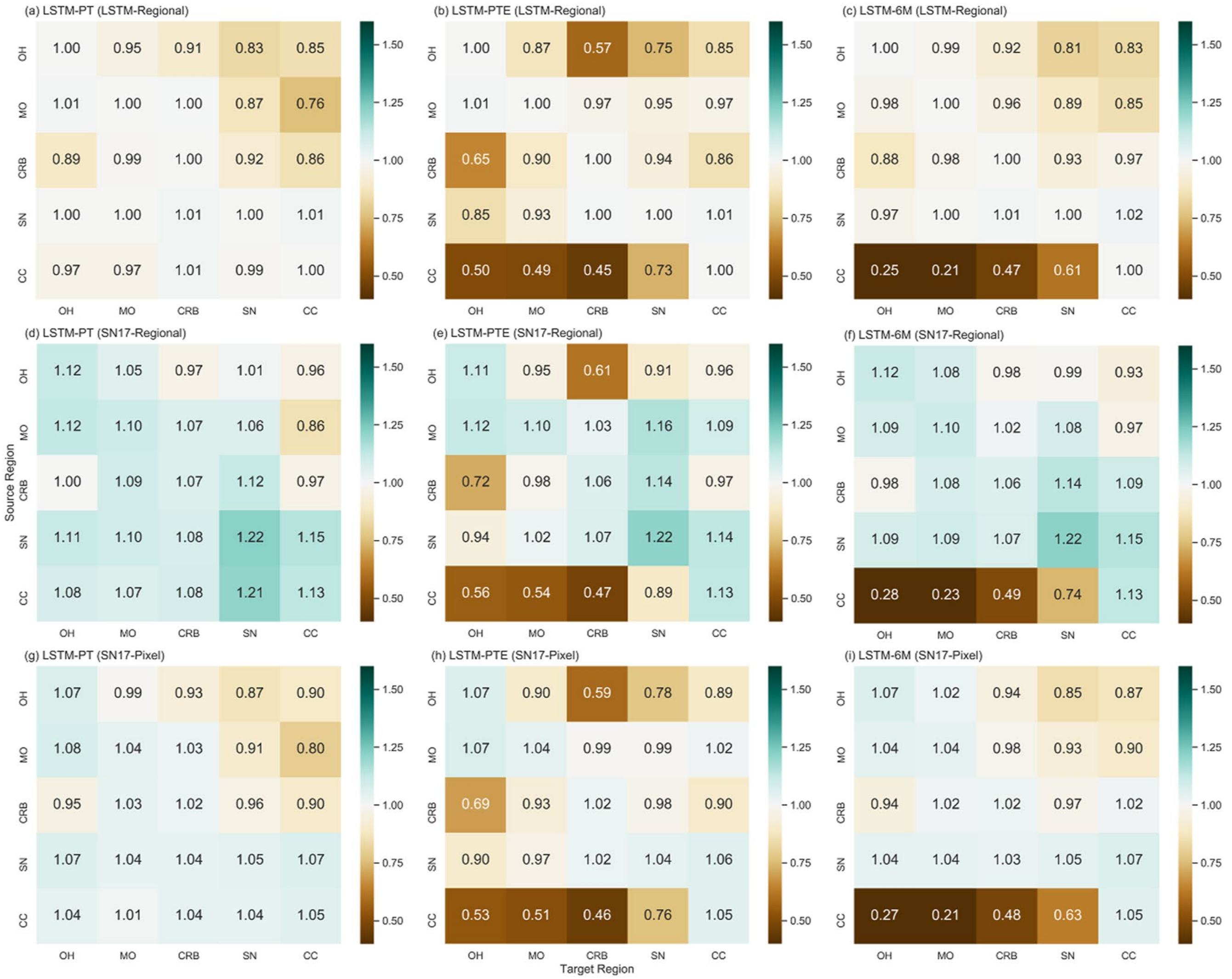

Figure 10. Results of the transfer learning experiments. In the top row, the transferred Long Short-Term Memory (LSTM) networks are compared to their local-region trained counterparts. In the middle row, the transferred LSTM networks are compared to the corresponding local-region-trained SN17 models. In the bottom row, the transferred LSTM networks are compared to the corresponding local-pixel-trained SN17 models. Values larger than 1.0 indicate good relative performance of the transferred LSTM models. OH = Ohio; MO = Missouri; CRB = Colorado River Basin; SN = Sierra Nevada; CC = Cascades.

to outperform the regionally or pixel-trained SN17 models. This suggests that future investigations may focus on how to improve the SN17 model by either incorporating more meteorological variables, or by enhancing the parameterized process representations within the model.

Finally, we note that the LSTM architecture exhibits better regional transferability than the SN17 model structure (Figure S10 in Supporting Information S1). This points to a fundamental difference between what is achieved when training an LSTM network as opposed to calibrating the SN17 model, where the former corresponds more closely to a structure learning problem (Gharari et al., 2021), while the latter is restricted to only parameter learning given a predefined model structure.

4.3.3. Remarks on Spatial Proximity Assumption for Region Delineation

Finally, we note that the success of network transferability seems to be related to the spatial proximity of the source and target regions. From Figure 10, we see that TL networks tend to transfer well only to the nearest adjacent regions. For example, the TL networks that use only precipitation and temperature (PT in Figure 10a), transfer well from *MO* to *OH* and *CRB*, and from *SN* to *CRB* and *CC*. As an exception, the same PT network structure transfers well from *CC* to CRB even though they are not geographically adjacent to each other. In general, however, one might speculate that the traditional method for region delineation may not be optimal from the point of view of knowledge transferability. Future study could focus on the use of other approaches for grouping pixels, based on important climatologic characteristics such as aridity, seasonality and fraction of precipitation falling as snow (Knoben et al., 2018), and seasonal precipitation and temperature patterns (Beck et al., 2018), or by data-based clustering of pixels based on patterns within the available data.

5. Conclusions, Remarks, and Outlook

5.1. Conclusions

In this study, we have investigated the potential for continental-scale LSTM-based modeling of snow accumulation and melt dynamics at the 4-km pixel scale over the CONUS. We have further investigated whether regional differences, based on geographical proximity, can be exploited to result in improved model performance. We followed a hierarchical training strategy in which a general LSTM architecture was first learned by assuming that a single network could represent SWE dynamics across the entire CONUS, followed by regional fine-tuning. We also investigated the benefits of using different kinds of input information, beyond that required by the SN17 model used by the US National Weather Service.

Overall, our results indicate that a single LSTM network, trained using data sampled from across the entire CONUS can provide remarkably good performance, as assessed via a variety of metrics, and that further regional-scale fine-tuning of the network results in only marginal improvement. Of particular relevance to future attempts to improve process-based representations (e.g., to improve the structure of SN17) is that the most accurate and robust performance is achieved when the network can access a variety of meteorological information (precipitation, temperature, dew point temperature, vapor pressure deficit, longwave radiation, and shortwave radiation), indicating that precipitation, temperature, and local elevation are not, by themselves, sufficiently informative to model the variability of snow dynamics at the continental scale. Further, when this range of meteorological information is provided to the network, the local information provided by elevation becomes redundant.

Comparison of the *LSTM-PTE* network with the physical-conceptual temperature-index-based SN17 model (where both are provided the same input information) indicates that the gating-operation and cell-states architecture of the LSTM enables it to learn a better representation of snow accumulation and melt dynamics than is encoded by SN17, and that by doing so a single CONUS-wide LSTM can significantly outperform an implementation of SN17 that is locally calibrated to each pixel. This result continues to hold even when regionally trained LSTMs are tested for regional transferability, suggesting considerable potential for improving physical-based representations to be applied CONUS-wide at the pixel resolution. In this context, LSTM-based modeling can serve as a valuable data compression tool that can assist the process of scientific hypothesis testing (Nearing et al., 2020), by providing insights regarding what kinds of information may be missing from existing process-based representations.

Of course, the data-intensive nature of LSTM-based modeling poses a potential barrier to the application of such techniques to data-scarce parts of the world where real-world meteorological forcing and SWE data are not widely available or have only limited temporal coverage. However, one reason for our sequential experimental design (proceeding from generic/global to specific/regional) was to explore the extent to which the use of a "pre-trained" LSTM network might be a reasonable way to circumvent the need for large amount of "local" training data (see also Kratzert et al., 2018). Our results indicate that such a strategy may indeed be viable, and future work should continue to explore to what specific/local extent this strategy can be pursued. In particular, it could be useful to investigate the smallest homogenous-local areal extents that can be differentiated while continuing to realize robust performance improvements. In this regard, studies will also need to be done regarding the minimum number of pixels for which data must be provided to efficiently achieve stable versions of trained CONUS-wide, Regional, and Local LSTM networks, and to assess what factors must be considered when designing a robust stratified sampling strategy for selecting representative pixels to ensure maximally informative data sets for training, evaluation and testing. This latter will need to consider snow-process-relevant diversity in terms of local ancillary variables related to various properties such as topographic and vegetation (Broxton et al., 2020).

In this study, a fixed time lag of 242 preceding days, along with the current days' information, is used to calculate the hidden state and cell input when updating the "state-dependent" parameters for each gating operation over the training period. This means that the cell state is initialized at zero at the beginning of every batch. This assumption regarding the nature of the average snowy season, when applied to the LSTM network, is shown to be capable of resulting in overall better performance than achieved by the various SN17 models. Further, the results of the transfer learning experiments suggest the existence of a tradeoff between model transferability and model complexity. In practice, one may have to individually optimize the sequence lengths for the CONUS-wide and regional-wide LSTMs. Future work should explore how changes to sequence length and the approach to initializing the cell state would affect the network performance and its impact on the results of transferability when different regionalization hypotheses are applied (Section 4.3.3).

5.2. Remarks on Model Benchmarks

Here, we have demonstrated only that a single ML algorithm (LSTM) can provide better performance than a single physical-conceptual temperature-index-based algorithm (SN17). While this is a good start, it clearly leaves many questions unasked and unanswered. In particular, we have not yet conducted a comparison with a variety of physically/process-based models—to cleanly perform such a comparison is nontrivial (Kratzert, Klotz, Shalev, et al., 2019; Lees et al., 2021) since different models may use different input information. However, this is certainly something that should be explored in future work. It would be helpful to investigate the performance of the LSTM networks in relation to other state-of-the-art models regarding whether the bias obtained is similar. By doing so, we can expect a high potential for gaining deeper insights into the relative strengths of data-based and physics-based approaches.

We note that a problem when comparing "physically based" models against data-based ones is that the former is typically constrained by conservation principles to limit the amount of SWE accumulation in a day to be less than or equal to the incoming precipitation. Precipitation undercatch encoded in the data, can be a source of bias that affects the comparison. Under such circumstances, a physically based model can be expected to consistently simulate lower values for snow accumulation, whereas a data-based approach that is restricted by mass balance constraints may be able to produce a better-quality simulation (Hoedt et al., 2021). In this regard, when the underlying data used is not internally consistent and adequate data preprocessing does not occur to remove biases from the data, data-based methods can have a real advantage.

5.3. Outlook

We expect that LSTM-based modeling of snow dynamics can be used to learn a universal model structure by leveraging the commonalities of meteorological data at various spatial locations and resolutions, thereby providing benefits in terms of hydrological modeling for data-scarce regions (Ma et al., 2021). Our study suggests that our LSTM-based strategy has the potential to be expanded to the development of continental and even global-scale systems for forecasting snow dynamics. In such systems, uncertainty quantification can be achieved either by applying Monte Carlo dropout (Fang et al., 2020; Klotz et al., 2021) or the use of multiple ML-based algorithms (Fleming & Goodbody, 2019). Given the large amount of data that is potentially available, further rigorous testing of the LSTM-based approach at pixel-scale resolution should be performed in both space and time (Gupta et al., 2014) with an emphasis on simulation performance with regard to various snow signatures including April 1st SWE and snow residence time (Lute & Luce, 2017; Zeng et al., 2018). Further, depth functions can be used to understand whether any given simulation lies more at the edge, the middle or even outside the clusters represented by the training data in a multi-variate sense, which may help with being able to understand the transferability of the models in a more general sense than examined in this paper (Chebana & Ouarda, 2011).

It also seems reasonable that combining LSTMs with other ML techniques, such as Convolutional Neural Networks (CNNs; LeCun et al., 1989), may be able to better exploit both the spatial and temporal information in the data set (Ehsani, Zarei, et al., 2021; Shi et al., 2015). Taking spatial information into account could help in better understanding snow dynamics and the relevance of each input variable, as the CNN layer-wise relevance propagation method could be used to produce heat maps that highlight areas and variables that are more relevant in order to produce certain SWE patterns (Bennett & Nijssen, 2021b).

Finally, physical explainability of ML-based results is a central contemporary challenge, one that is key to widespread acceptance of Artificial Intelligence (AI). So far, the success of ML has not been translated into significantly improved knowledge of the processes underlying snow dynamics. More efforts should be made to tackling this issue in a hydrologic (Fleming, Vesselinov, & Goodbody, 2021) and meteorological context (McGovern et al., 2019). In fact, this can be advanced by symbiotic integration of physically based and data-based models that facilitate the development of physics-AI synergy modeling approaches (Reichstein et al., 2019). Recent attempts have included replacing internal process equations with networks that have the ability to learn from data (Bennett & Nijssen, 2021a), the embedding of physically based representations into ML networks (Jiang et al., 2020), and the imposition of mass balance constraints into ML (Hoedt et al., 2021; Nearing et al., 2021).

In our view, another potential approach is to use symbolic regression to facilitate the development of hybrid modeling systems that can learn "physically understandable" process representations (Udrescu & Tegmark, 2020) while adhering to the principle of parsimony (Occam's Razor; see discussion by Weijs & Ruddell, 2020) thereby turning the pure "black-box" model into the opaque "glass-box" model (Rai, 2020). One of the directions that we intend to pursue is to automate the search for physically consistent parameter transfer functions by a process of learning from large data sets (Feigl et al., 2020; Gharari et al., 2021; Klotz et al., 2017). Notice however that pursuing this direction may be subject to an "accuracy-interpretability" dilemma that has been suggested to arise from a perhaps irreconcilable conflict between a model's predictive accuracy and the possibility of understanding its behaviors (Florez-Lopez & Ramon-Jeronimo, 2015). It would be valuable to devote more attention to investigating the use of DL interpretation methods to facilitate AI-assisted scientific discovery (Jiang et al., 2022). By doing so, the advances in AI might more properly be able to leverage existing theories, helping to revolutionize the next generation of earth and environmental sciences (ESS) models in the near future (Fleming, Watson, et al., 2021).

Data Availability Statement

The PRISM daily 4-km temperature and precipitation data are available at http://www.prism.oregonstate.edu/. The NLDAS-2 data are available online (http://www.emc.ncep.noaa.gov/mmb/nldas/). The University of Arizona snow data are available at the NSIDC data center (https://doi.org/10.5067/0GGPB220EX6A).

Acknowledgments
This work was partially supported by grants NA18OAR4590397 from the National Oceanic and Atmospheric Administration (NOAA) OAR's OWAQ. The second author acknowledges partial support by the Australian Centre of Excellence for Climate System Science (CE110001028). The authors would like to thank the WRR editorial team including Charles Luce (Editor), the Associate Editor, reviewer Thorsten Wagner, and two anonymous reviewers for taking their time to provide constructive comments. Further thanks are due to Sungwook Wi and Tirthankar Roy for providing the SNOW17 code and Patrick Broxton for discussions regarding the data used, and to Andrew Bennett, Luis De La Fuente and Mohammad Reza Ehsani for pre-reviewing the early version of manuscript.

References

Ahmed, E., & Rothenberger, A. (2015). KFeSbTe$_3$: A quaternary chalcogenide aerogel for preferential adsorption of polarizable hydrocarbons and gases. *Journal of Materials Chemistry A*, *3*, 7786–7792. https://doi.org/10.1039/c4ta06957a

Anderson, E. A. (1973). *National Weather Service river forecast system: Snow accumulation and ablation model* (Vol. 17). US Department of Commerce, National Oceanic and Atmospheric Administration, National Weather Service.

Anderson, E. A. (2006). *Snow accumulation and ablation model–SNOW-17* (p. 61). US National Weather Service.

Arevalo, J., Welty, J., Fan, Y., & Zeng, X. (2021). Implementation of snowpack treatment in the CPC water balance model and its impact on drought assessment. *Journal of Hydrometeorology*. https://doi.org/10.1175/JHM%2DD%2D20%2D0201.1

Bales, R. C., Molotch, N. P., Painter, T. H., Dettinger, M. D., Rice, R., & Dozier, J. (2006). Mountain hydrology of the western United States. *Water Resources Research*, *42*(8). https://doi.org/10.1029/2005WR004387

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), 303–309. https://doi.org/10.1038/nature04141

Barrett, A. P. (2003). *National operational hydrologic remote sensing center snow data assimilation system (SNODAS) products at NSIDC* (p. 19). National Snow and Ice Data Center, Cooperative Institute for Research in Environmental Sciences.

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

Beck, H. E., Zimmermann, N. E., McVicar, T. R., Vergopolan, N., Berg, A., & Wood, E. F. (2018). Present and future Köppen-Geiger climate classification maps at 1-km resolution. *Scientific Data*, *5*(1), 1–12. https://doi.org/10.1038/sdata.2018.214

Bengio, Y. (2012). Practical recommendations for gradient-based training of deep architectures. In *Neural networks: Tricks of the trade* (pp. 437–478). Springer. https://doi.org/10.1007/978-3-642-35289-8_26

Bennett, A., & Nijssen, B. (2021a). Deep learned process parameterizations provide better representations of turbulent heat fluxes in hydrologic models. *Water Resources Research*, *57*(5), e2020WR029328. https://doi.org/10.1029/2020wr029328

Bennett, A., & Nijssen, B. (2021b). Explainable AI uncovers how neural networks learn to regionalize in simulations of turbulent heat fluxes at FluxNet sites. *Earth and Space Science Open Archive ESSOAr*. https://doi.org/10.1002/essoar.10506880.1

Boisvenue, C., & Running, S. W. (2006). Impacts of climate change on natural forest productivity–evidence since the middle of the 20th century. *Global Change Biology*, *12*(5), 862–882. https://doi.org/10.1111/j.1365-2486.2006.01134.x

Broxton, P. D., Dawson, N., & Zeng, X. (2016). Linking snowfall and snow accumulation to generate spatial maps of SWE and snow depth. *Earth and Space Science*, *3*(6), 246–256. https://doi.org/10.1002/2016EA000174

Broxton, P. D., van Leeuwen, W., & Biederman, J. A. (2017). December. SWANN: The snow water artificial neural network modelling system. *AGU Fall Meeting Abstracts*, *2017*, C43B.

Broxton, P. D., Van Leeuwen, W. J., & Biederman, J. A. (2019). Improving snow water equivalent maps with machine learning of snow survey and lidar measurements. *Water Resources Research*, *55*(5), 3739–3757. https://doi.org/10.1029/2018WR024146

Broxton, P. D., van Leeuwen, W. J., & Biederman, J. A. (2020). Forest cover and topography regulate the thin, ephemeral snowpacks of the semiarid Southwest United States. *Ecohydrology*, *13*(4), e2202. https://doi.org/10.1002/eco.2202

Broxton, P. D., Zeng, X., & Dawson, N. (2016). Why do global reanalyses and land data assimilation products underestimate snow water equivalent? *Journal of Hydrometeorology*, *17*(11), 2743–2761. https://doi.org/10.1175/JHM-D-16-0056.1

Broxton, P. D., Zeng, X., Sulla-Menashe, D., & Troch, P. A. (2014). A global land cover climatology using MODIS data. *Journal of Applied Meteorology and Climatology*, *53*(6), 1593–1605. https://doi.org/10.1175/jamc-d-13-0270.1

Brun, E., David, P., Sudul, M., & Brunot, G. (1992). A numerical model to simulate snow-cover stratigraphy for operational avalanche forecasting. *Journal of Glaciology*, *38*(128), 13–22. https://doi.org/10.1017/s0022143000009552

Buckingham, D., Skalka, C., & Bongard, J. (2015). Inductive machine learning for improved estimation of catchment-scale snow water equivalent. *Journal of Hydrology*, *524*, 311–325. https://doi.org/10.1016/j.jhydrol.2015.02.042

Chaney, N. W., Metcalfe, P., & Wood, E. F. (2016). HydroBlocks: A field-scale resolving land surface model for application over continental extents. *Hydrological Processes*, *30*(20), 3543–3559. https://doi.org/10.1002/hyp.10891

Chebana, F., & Ouarda, T. B. (2011). Depth-based multivariate descriptive statistics with hydrological applications. *Journal of Geophysical Research: Atmospheres*, *116*(D10). https://doi.org/10.1029/2010jd015338

Cho, E., & Jacobs, J. M. (2020). Extreme value snow water equivalent and snowmelt for infrastructure design over the contiguous United States. *Water Resources Research*, *56*(10), e2020WR028126. https://doi.org/10.1029/2020WR028126

Cho, E., Jacobs, J. M., & Vuyovich, C. M. (2020). The value of long-term (40 years) airborne gamma radiation SWE record for evaluating three observation-based gridded SWE data sets by seasonal snow and land cover classifications. *Water Resources Research*, *56*(1). https://doi.org/10.1029/2019WR025813

Christensen, N. S., Wood, A. W., Voisin, N., Lettenmaier, D. P., & Palmer, R. N. (2004). The effects of climate change on the hydrology and water resources of the Colorado River basin. *Climatic Change*, *62*(1), 337–363. https://doi.org/10.1023/B:CLIM.0000013684.13621.1f

Clark, M. P., Kavetski, D., & Fenicia, F. (2011). Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resources Research*, *47*(9). https://doi.org/10.1029/2010WR009827

Clark, M. P., Schaefli, B., Schymanski, S. J., Samaniego, L., Luce, C. H., Jackson, B. M., et al. (2016). Improving the theoretical underpinnings of process-based hydrologic models. *Water Resources Research*, *52*(3), 2350–2365. https://doi.org/10.1002/2015WR017910

Czyzowska-Wisniewski, E. H., van Leeuwen, W. J., Hirschboeck, K. K., Marsh, S. E., & Wisniewski, W. T. (2015). Fractional snow cover estimation in complex alpine-forested environments using an artificial neural network. *Remote Sensing of Environment*, *156*, 403–417. https://doi.org/10.1016/j.rse.2014.09.026

Daly, C., Neilson, R. P., & Phillips, D. L. (1994). A statistical-topographic model for mapping climatological precipitation over mountainous terrain. *Journal of Applied Meteorology and Climatology*, *33*(2), 140–158. https://doi.org/10.1175/1520-0450(1994)033<0140:astmfm>2.0.co;2;2

Dawson, N., Broxton, P., & Zeng, X. (2018). Evaluation of remotely sensed snow water equivalent and snow cover extent over the contiguous United States. *Journal of Hydrometeorology*, *19*(11), 1777–1791. https://doi.org/10.1175/JHM-D-18-0007.1

Deems, J. S., Painter, T. H., Barsugli, J. J., Belnap, J., & Udall, B. (2013). Combined impacts of current and future dust deposition and regional warming on Colorado River Basin snow dynamics and hydrology. *Hydrology and Earth System Sciences*, *17*(11), 4401–4413. https://doi.org/10.5194/hess-17-4401-2013

Duan, Q., Sorooshian, S., & Gupta, V. (1992). Effective and efficient global optimization for conceptual rainfall-runoff models. *Water Resources Research*, *28*(4), 1015–1031. https://doi.org/10.1029/91WR02985

Ehsani, M. R., Behrangi, A., Adhikari, A., Song, Y., Huffman, G. J., Adler, R. F., et al. (2021). Assessment of the advanced very high-resolution radiometer (AVHRR) for snowfall retrieval in high latitudes using CloudSat and machine learning. *Journal of Hydrometeorology*. https://doi.org/10.1175/jhm-d-20-0240.1

Ehsani, M. R., Zarei, A., Gupta, H. V., Barnard, K., & Behrangi, A. (2021). *Nowcasting-Nets: Deep neural network structures for precipitation nowcasting using IMERG*. Retrieved from https://arxiv.org/abs/2108.06868

Fang, K., Kifer, D., Lawson, K., & Shen, C. (2020). Evaluating the potential and challenges of an uncertainty quantification method for long short-term memory models for soil moisture predictions. *Water Resources Research*, *56*(12), e2020WR028095. https://doi.org/10.1029/2020WR028095

Feigl, M., Herrnegger, M., Klotz, D., & Schulz, K. (2020). Function space optimization: A symbolic regression method for estimating parameter transfer functions for hydrological models. *Water Resources Research*, *56*(10), e2020WR027385. https://doi.org/10.1029/2020wr027385

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

Fleming, S. W., Watson, J. R., Ellenson, A., Cannon, A. J., & Vesselinov, V. C. (2021). Machine learning in Earth and environmental science requires education and research policy reforms. *Nature Geoscience*, *14*(12), 878–880. https://doi.org/10.1038/s41561-021-00865-3

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

Florez-Lopez, R., & Ramon-Jeronimo, J. M. (2015). Enhancing accuracy and interpretability of ensemble strategies in credit risk assessment. A correlated-adjusted decision forest proposal. *Expert Systems with Applications*, *42*(13), 5737–5753. https://doi.org/10.1016/j.eswa.2015.02.042

Ford, C. M., Kendall, A. D., & Hyndman, D. W. (2020). Effects of shifting snowmelt regimes on the hydrology of non-alpine temperate landscapes. *Journal of Hydrology*, *590*, 125517. https://doi.org/10.1016/j.jhydrol.2020.125517

Garvelmann, J., Pohl, S., & Weiler, M. (2015). Spatio-temporal controls of snowmelt and runoff generation during rain-on-snow events in a mid-latitude mountain catchment. *Hydrological Processes*, *29*(17), 3649–3664. https://doi.org/10.1002/hyp.10460

Gharaei-Manesh, S., Fathzadeh, A., & Taghizadeh-Mehrjardi, R. (2016). Comparison of artificial neural network and decision tree models in estimating spatial distribution of snow depth in a semi-arid region of Iran. *Cold Regions Science and Technology*, *122*, 26–35. https://doi.org/10.1016/j.coldregions.2015.11.004

Gharari, S., Gupta, H. V., Clark, M. P., Hrachowitz, M., Fenicia, F., Matgen, P., & Savenije, H. H. (2021). Understanding the information content in the hierarchy of model development decisions: Learning from data. *Water Research*, *57*, e2020WR027948. https://doi.org/10.1029/2020WR027948

Gong, W., Gupta, H. V., Yang, D., Sricharan, K., & Hero, A. O., III (2013). Estimating epistemic and aleatory uncertainties during hydrologic modeling: An information theoretic approach. *Water Resources Research*, *49*(4), 2253–2273. https://doi.org/10.1002/wrcr.20161

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., Perrin, C., Blöschl, G., Montanari, A., Kumar, R., Clark, M., & Andréassian, V. (2014). Large-sample hydrology: A need to balance depth with breadth. *Hydrology and Earth System Sciences*, *18*(2), 463–477. https://doi.org/10.5194/hess-18-463-2014

He, M., Hogue, T. S., Franz, K. J., Margulis, S. A., & Vrugt, J. A. (2011a). Characterizing parameter sensitivity and uncertainty for a snow model across hydroclimatic regimes. *Advances in Water Resources*, *34*(1), 114–127. https://doi.org/10.1016/j.advwatres.2010.10.002

He, M., Hogue, T. S., Franz, K. J., Margulis, S. A., & Vrugt, J. A. (2011b). Corruption of parameter behavior and regionalization by model and forcing data errors: A Bayesian example using the SNOW17 model. *Water Resources Research*, *47*(7). https://doi.org/10.1029/2010WR009753

Henn, B., Newman, A. J., Livneh, B., Daly, C., & Lundquist, J. D. (2018). An assessment of differences in gridded precipitation datasets in complex terrain. *Journal of Hydrology*, *556*, 1205–1219. https://doi.org/10.1016/j.jhydrol.2017.03.008

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

Hoedt, P. J., Kratzert, F., Klotz, D., Halmich, C., Holzleitner, M., Nearing, G., et al. (2021). *MC-LSTM: Mass-Conserving LSTM*. arXiv preprint arXiv:2101.05186.

Hrachowitz, M., Savenije, H. H. G., Blöschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J. W., et al. (2013). A decade of predictions in ungauged basins (PUB)—A review. *Hydrological Sciences Journal*, *58*(6), 1198–1255. https://doi.org/10.1080/02626667.2013.803183

Huo, X., Gupta, H., Niu, G. Y., Gong, W., & Duan, Q. (2019). Parameter sensitivity analysis for computationally intensive spatially distributed dynamical environmental systems models. *Journal of Advances in Modeling Earth Systems*, *11*(9), 2896–2909. https://doi.org/10.1029/2018MS001573

Jarvis, A. (2008). *Hole-field seamless SRTM data*. International Centre for Tropical Agriculture (CIAT). Retrieved from http://srtm.csi.cgiar.org

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

Jiang, S., Zheng, Y., Wang, C., & Babovic, V. (2022). Uncovering flooding mechanisms across the contiguous United States through interpretive deep learning on representative catchments. *Water Resources Research*, *58*, e2021WR030185. https://doi.org/10.1029/2021WR030185

Jin, J., Gao, X., Sorooshian, S., Yang, Z. L., Bales, R., Dickinson, R. E., et al. (1999). One-dimensional snow water and energy balance model for vegetated surfaces. *Hydrological Processes*, *13*(14-15), 2467–2482. https://doi.org/10.1002/(SICI)1099-1085(199910)13:14/15<2467::AID-HYP861>3.0.CO;2-J

Khatami, S., Peterson, T. J., Peel, M. C., & Western, A. (2020). *Evaluating catchment models as multiple working hypotheses: On the role of error metrics, parameter sampling, model structure, and data information content*. arXiv preprint arXiv:2009.00729.

Kingma, D. P. & Ba, J. (2014). *Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980*.

Klotz, D., Herrnegger, M., & Schulz, K. (2017). Symbolic regression for the estimation of transfer functions of hydrological models. *Water Resources Research*, *53*(11), 9402–9423. https://doi.org/10.1002/2017wr021253

Klotz, D., Kratzert, F., Gauch, M., Keefe Sampson, A., Brandstetter, J., Klambauer, G., et al. (2021). Uncertainty estimation with deep learning for rainfall–runoff modelling. *Hydrology and Earth System Sciences Discussions*, 1–32. https://doi.org/10.5194/hess-2021-154

Knoben, W. J., Woods, R. A., & Freer, J. E. (2018). A quantitative hydrological climate classification evaluated with independent streamflow data. *Water Resources Research*, *54*(7), 5088–5109. https://doi.org/10.1029/2018WR022913

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. *Water Resources Research*, *55*(12), 11344–11354. https://doi.org/10.1029/2019WR026065

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, *23*(12), 5089–5110. https://doi.org/10.5194/hess-23-5089-2019

Kuter, S. (2021). Completing the machine learning saga in fractional snow cover estimation from MODIS Terra reflectance data: Random forests versus support vector regression. *Remote Sensing of Environment*, *255*, 112294. https://doi.org/10.1016/j.rse.2021.112294

Kuter, S., Akyurek, Z., & Weber, G. W. (2018). Retrieval of fractional snow covered area from MODIS data by multivariate adaptive regression splines. *Remote Sensing of Environment*, *205*, 236–252. https://doi.org/10.1016/j.rse.2017.11.021

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, *1*(4), 541–551. https://doi.org/10.1162/neco.1989.1.4.541

LeCun, Y. A., Bottou, L., Orr, G. B., & Müller, K. R. (2012). Efficient backprop. In *Neural networks: Tricks of the trade* (pp. 9–48). Springer. https://doi.org/10.1007/978-3-642-35289-8_3

Lees, T., Buechel, M., Anderson, B., Slater, L., Reece, S., Coxon, G., & Dadson, S. J. (2021). Benchmarking data-driven rainfall–runoff models in Great Britain: A comparison of long short-term memory (LSTM)-based models with four lumped conceptual models. *Hydrology and Earth System Sciences*, *25*(10), 5517–5534. https://doi.org/10.5194/hess-25-5517-2021

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*, *44*(12), 6163–6172. https://doi.org/10.1002/2017GL073551

Liang, X., Lettenmaier, D. P., Wood, E. F., & Burges, S. J. (1994). A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *Journal of Geophysical Research*, *99*(D7), 14415–14428. https://doi.org/10.1029/94JD00483

Lute, A. C., & Luce, C. H. (2017). Are model transferability and complexity antithetical? Insights from validation of a variable-complexity empirical snow model in space and time. *Water Resources Research*, *53*(11), 8825–8850. https://doi.org/10.1002/2017WR020752

Ma, K., Feng, D., Lawson, K., Tsai, W. P., Liang, C., Huang, X., et al. (2021). Transferring hydrologic data across continents–leveraging data-rich regions to improve hydrologic prediction in data-sparse regions. *Water Resources Research*, *57*(5), e2020WR028600. https://doi.org/10.1029/2020WR028600

Mankin, J. S., Viviroli, D., Singh, D., Hoekstra, A. Y., & Diffenbaugh, N. S. (2015). The potential for snow to supply human water demand in the present and future. *Environmental Research Letters*, *10*(11), 114016. https://doi.org/10.1088/1748-9326/10/11/114016

Marks,D.,Domingo,J.,Susong,D.,Link,T.,&Garen,D.(1999).Aspatiallydistributedenergybalancesnowmeltmodelforapplicationinmountainbasins. *Hydrological Processes*, *13*(12-13), 1935–1959. https://doi.org/10.1002/(sici)1099-1085(199909)13:12/13<1935::aid-hyp868>3.0.co;2-c

McGovern, A., Lagerquist, R., Gagne, D. J., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., & Smith, T. (2019). Making the black box more transparent: Understanding the physical implications of machine learning. *Bulletin of the American Meteorological Society*, *100*(11), 2175–2199. https://doi.org/10.1175/bams-d-18-0195.1

Mote, P. W. (2006). Climate-driven variability and trends in mountain snowpack in western North America. *Journal of Climate*, *19*(23), 6209–6220. https://doi.org/10.1175/JCLI3971.1

Musselman, K. N., Addor, N., Vano, J. A., & Molotch, N. P. (2021). Winter melt trends portend widespread declines in snow water resources. *Nature Climate Change*, *11*, 1–424. https://doi.org/10.1038/s41558-021-01014-9

Musselman, K. N., Lehner, F., Ikeda, K., Clark, M. P., Prein, A. F., Liu, C., et al. (2018). Projected increases and shifts in rain-on-snow flood risk over western North America. *Nature Climate Change*, *8*(9), 808–812. https://doi.org/10.1038/s41558-018-0236-4

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I—A discussion of principles. *Journal of Hydrology*, *10*(3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

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

Nearing, G. S., Ruddell, B. L., Bennett, A. R., Prieto, C., & Gupta, H. V. (2020). Does information theory provide a new paradigm for earth science? Hypothesis testing. *Water Resources Research*, *56*(2). https://doi.org/10.1029/2019WR024918

Nijssen, B., O'Donnell, G. M., Hamlet, A. F., & Lettenmaier, D. P. (2001). Hydrologic sensitivity of global rivers to climate change. *Climatic Change*, *50*(1), 143–175. https://doi.org/10.1023/A:1010616428763

Niu, G. Y., & Yang, Z. L. (2007). An observation-based formulation of snow cover fraction and its evaluation over large North American river basins. *Journal of geophysical research*, *112*(D21). https://doi.org/10.1029/2007JD008674

Niu, G. Y., Yang, Z. L., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M., et al. (2011). The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements. *Journal of Geophysical Research: Atmospheres*, *116*(D12). https://doi.org/10.1029/2010jd015139

Ntokas, K. F., Odry, J., Boucher, M. A., & Garnaud, C. (2021). Investigating ANN architectures and training to estimate snow water equivalent from snow depth. *Hydrology and Earth System Sciences*, *25*(6), 3017–3040. https://doi.org/10.5194/hess-25-3017-2021

Odry, J., Boucher, M. A., Cantet, P., Lachance-Cloutier, S., Turcotte, R., & St-Louis, P. Y. (2020). Using artificial neural networks to estimate snow water equivalent from snow depth. *Canadian Water Resources Journal/Revue canadienne des ressources hydriques*, *45*(3), 252–268. https://doi.org/10.1080/07011784.2020.1796817

Pohl, S., Marsh, P., & Liston, G. E. (2006). Spatial-temporal variability in turbulent fluxes during spring snowmelt. *Arctic Antarctic and Alpine Research*, *38*(1), 136–146. https://doi.org/10.1657/1523-0430%282006%29038%5B0136:SVITFD%5D2.0.CO;2

Pokhrel, P., Gupta, H. V., & Wagener, T. (2008). A spatial regularization approach to parameter estimation for a distributed watershed model. *Water Resources Research*, *44*(12). https://doi.org/10.1029/2007WR006615

Qin, Y., Abatzoglou, J. T., Siebert, S., Huning, L. S., AghaKouchak, A., Mankin, J. S., et al. (2020). Agricultural risks from changing snowmelt. *Nature Climate Change*, *10*(5), 459–465. https://doi.org/10.1038/s41558-020-0746-8

Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science*, *48*(1), 137–141. https://doi.org/10.1007/s11747-019-00710-5

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

Revuelto, J., Billecocq, P., Tuzet, F., Cluzet, B., Lamare, M., Larue, F., & Dumont, M. (2020). Random forests as a tool to understand the snow depth distribution and its evolution in mountain areas. *Hydrological Processes*. https://doi.org/10.1002/hyp.13951

Robinson, D. A., Dewey, K. F., & Heim, R. R., Jr (1993). Global snow cover monitoring: An update. *Bulletin of the American Meteorological Society*, *74*(9), 1689–1696. https://doi.org/10.1175/1520-0477(1993)074<1689:GSCMAU>2.0.CO;2

Samaniego, L., Kumar, R., & Attinger, S. (2010). Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. *Water Resources Research*, *46*(5). https://doi.org/10.1029/2008WR007327

Schaefli, B., & Gupta, H. V. (2007). Do Nash values have value? *Hydrological Processes*, *21*, 2075–2080. https://doi.org/10.1002/hyp.6825

Schaefli, B., Hingray, B., Niggli, M., & Musy, A. (2005). A conceptual glacio-hydrological model for high mountainous catchments. *Hydrology and Earth System Sciences*, *9*(1/2), 95–109. https://doi.org/10.5194/hess-9-95-2005

Schoups, G., Van de Giesen, N. C., & Savenije, H. H. G. (2008). Model complexity control for hydrologic prediction. *Water Resources Research*, *44*(12). https://doi.org/10.1029/2008wr006836

Shen, C. (2018). A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resources Research*, *54*(11), 8558–8593. https://doi.org/10.1029/2018WR022643

Shen, C., Laloy, E., Elshorbagy, A., Albert, A., Bales, J., Chang, F. J., et al. (2018). HESS Opinions: Incubating deep-learning-powered hydrologic science advances as a community. *Hydrology and Earth System Sciences*, *22*(11), 5639–5656. https://doi.org/10.5194/hess-22-5639-2018

Shi, X., Chen, Z., Wang, H., Yeung, D., Wong, W., & Woo, W. (2015). *Convolutional LSTM network: A machine learning approach for precipitation nowcasting*. Retrieved from https://arxiv.org/abs/1506.04214

Shindell, D., Kuylenstierna, J. C., Vignati, E., van Dingenen, R., Amann, M., Klimont, Z., et al. (2012). Simultaneously mitigating near-term climate change and improving human health and food security. *Science*, *335*(6065), 183–189. https://doi.org/10.1126/science.1210026

Simpkins, G. (2018). Snow-related water woes. *Nature Climate Change*, *8*(11), 945. https://doi.org/10.1038/s41558-018-0330-7

Sivapalan, M., Takeuchi, K., Franks, S. W., Gupta, V. K., Karambiri, H., Lakshmi, V., et al. (2003). IAHS decade on predictions in ungauged basins (PUB), 2003–2012: Shaping an exciting future for the hydrological sciences. *Hydrological Sciences Journal*, *48*(6), 857–880. https://doi.org/10.1623/hysj.48.6.857.51421

Snauffer, A. M., Hsieh, W. W., Cannon, A. J., & Schnorbus, M. A. (2018). Improving gridded snow water equivalent products in British Columbia, Canada: Multi-source data fusion by neural network models. *The Cryosphere*, *12*(3), 891–905. https://doi.org/10.5194/tc-12-891-2018

Strasser, U., Etchevers, P., & Lejeune, Y. (2002). Inter-comparison of two snow models with different complexity using data from an alpine site: Selected paper from EGS General Assembly, Nice, April-2000 (symposium OA36). *Hydrology Research*, *33*(1), 15–26. https://doi.org/10.2166/nh.2002.0002

Swenson, S. C., & Lawrence, D. M. (2012). A new fractional snow-covered area parameterization for the Community Land Model and its effect on the surface energy balance. *Journal of Geophysical Research: Atmospheres*, *117*(D21). https://doi.org/10.1029/2012JD018178

Tabari, H., Marofi, S., Abyaneh, H. Z., & Sharifi, M. R. (2010). Comparison of artificial neural network and combined models in estimating spatial distribution of snow depth and snow water equivalent in Samsami basin of Iran. *Neural Computing & Applications*, *19*(4), 625–635. https://doi.org/10.1007/s00521-009-0320-9

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

Tribbeck, M. J., Gurney, R. J., Morris, E. M., & Pearson, D. W. C. (2004). A new snow-SVAT to simulate the accumulation and ablation of seasonal snow cover beneath a forest canopy. *Journal of Glaciology*, *50*, 171–182. https://doi.org/10.3189/172756504781830187

Udrescu, S. M., & Tegmark, M. (2020). AI Feynman: A physics-inspired method for symbolic regression. *Science Advances*, *6*(16), eaay2631. https://doi.org/10.1126/sciadv.aay2631

Wang, Y. H., Broxton, P., Fang, Y., Behrangi, A., Barlage, M., Zeng, X., & Niu, G. Y. (2019). A wet-bulb temperature-based rain-snow partitioning scheme improves snowpack prediction over the drier western United States. *Geophysical Research Letters*, *46*(23), 13825–13835. https://doi.org/10.1029/2019GL085722

Weijs, S. V., & Ruddell, B. L. (2020). Debates: Does information theory provide a new paradigm for earth science? Sharper predictions using Occam's digital razor. *Water Resources Research*, *56*(2). https://doi.org/10.1029/2019WR026471

Welty, J., & Zeng, X. (2021). Characteristics and causes of extreme snowmelt over the conterminous United States. *Bulletin of the American Meteorological Society*, *102*, 1–E1542. https://doi.org/10.1175/BAMS-D-20-0182.1

Westerling, A. L. (2016). Increasing western US forest wildfire activity: Sensitivity to changes in the timing of spring. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1696), 20150178. https://doi.org/10.1098/rstb.2015.0178

Wunsch, A., Liesch, T., & Broda, S. (2021). Groundwater level forecasting with artificial neural networks: A comparison of long short-term memory (LSTM), convolutional neural networks (CNNs), and non-linear autoregressive networks with exogenous input (NARX). *Hydrology and Earth System Sciences*, *25*(3), 1671–1687. https://doi.org/10.5194/hess-25-1671-2021

Xia, Y., Ek, M., Wei, H., & Meng, J. (2012). Comparative analysis of relationships between NLDAS-2 forcings and model outputs. *Hydrological Processes*, *26*(3), 467–474. https://doi.org/10.1002/hyp.8240

Xiao, M. (2021). A warning of earlier snowmelt. *Nature Climate Change*, 1–2. https://doi.org/10.1038/s41558-021-01024-7

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). *How transferable are features in deep neural networks?*

Zeng, X., Broxton, P., & Dawson, N. (2018). Snowpack change from 1982 to 2016 over conterminous United States. *Geophysical Research Letters*, *45*(23), 12–940. https://doi.org/10.1029/2018GL079621