Using Temporal Deep Learning Models to Estimate Historical and End-Century Daily Snow Water Equivalent over the Rocky Mountains

Shiheng Duan[1], Paul Ullrich[2], Mark Risser[3], and Alan Rhoades[3]

[1]University of California, Davis
[2]University of California Davis
[3]Lawrence Berkeley National Laboratory

March 31, 2023

Using Temporal Deep Learning Models to Estimate Daily Snow Water Equivalent over the Rocky Mountains

Shiheng Duan[1], **Paul Ullrich**[1], **Mark Risser**[2], **Alan Rhoades**[2]

[1]Atmospheric Science Graduate Group, University of California, Davis
[2]Lawrence Berkeley National Laboratory

Key Points:

- Three different deep learning models are assessed for daily snow water equivalent prediction.
- Sensitivity tests provide evidence the DL models follow physical laws.
- Snow water equivalent fraction is used to alleviate problems with spatial extrapolation.

Corresponding author: Shiheng Duan, `shiduan@ucdavis.edu`

Abstract

In this study we construct and compare three different deep learning (DL) models for estimating daily snow water equivalent (SWE) from high-resolution gridded meteorological fields over the Rocky Mountain region. To train the DL models, Snow Telemetry (SNOTEL) station-based SWE observations are used as the prediction target. All DL models produce higher median Nash-Sutcliffe Efficiency (NSE) values than a process-based SWE model and products, although mean squared errors also tend to be higher. Sensitivity of the SWE prediction to the model's input variables is analyzed using an explainable artificial intelligence (XAI) method, yielding insight into the physical relationships learned by the models. This method reveals the dominant role precipitation and temperature play in snowpack dynamics. In applying our models to estimate SWE throughout the Rocky Mountains, an extrapolation problem arises since the statistical properties of SWE (e.g., annual maximum) and geographical properties of individual grid points (e.g., elevation) differ from the training data. This problem is solved by switching the prediction target to SWE fraction to alleviate extrapolation for all tested DL models. Our work shows that the DL models are promising tools for estimating SWE, and sufficiently capture relevant physical relationships to make them useful for spatial and temporal extrapolation of SWE values.

1 Introduction

Snowpack is a central component of the hydrologic cycle in montane regions, and its capacity to act as a reservoir for seasonal water storage is of vital importance to downstream communities. This is especially true for watersheds in mid-to-high latitudes and at altitudes where streamflow is derived largely from snowmelt (Berghuijs et al., 2019). Snow water equivalent (SWE), defined as the equivalent amount of liquid water stored in the snowpack if it were to be instantaneously melted, is the metric most commonly employed by water managers to estimate and evenly compare water content of the snowpack across regions. Climate change has already and will continue to significantly reduce both mean and maximum annual SWE, which will have repercussions for both streamflow and groundwater dynamics, and in turn pose major challenges on water managers (Rhoades, Ullrich, & Zarzycki, 2018; Livneh & Badger, 2020; X. Chen et al., 2021; Hatchett et al., 2022; Rhoades et al., 2022). However, estimating the exact magnitude, timing and persistence of SWE across various mountain ranges remains a scientific grand challenge (Siirila-Woodburn et al., 2021). Thus, there is considerable value for both science and society in the development of novel methods that can more precisely estimate spatiotemporally continuous SWE values over mountainous regions, both historically and into the future.

Substantial and rapid progress in the development of machine learning (ML) and deep learning (DL) methods, and corresponding hardware advancements related to graphical processing units (GPUs), has stimulated promising research in the use of ML and DL-based models for problems in Earth system science (Feng et al., 2022). ML models have also been employed and have proven valuable for estimating SWE, although the majority of this research has focused on historical SWE estimation from existing snow and snow-related datasets. For instance, Snauffer et al. (2018) used an artificial neural network (ANN) model to estimate SWE from several reanalysis products. Their ML-generated SWE estimation exhibited better agreement with station observations, compared to those derived from a Variable Infiltration Capacity (VIC) hydrological model simulation. Odry et al. (2020) and Ntokas et al. (2021) designed an ANN model to predict SWE and demonstrated that their ML model outperformed the benchmark regression model. Their input variables included snow depth, temperature, accumulated precipitation and several indices such as the number of snow-free days and the number of layers in the snowpack. Random forest methods have also been adopted to bias correct gridded SWE products (King et al., 2020).

To date, ML-based SWE estimation has largely relied on inference or emulation of existing snow-related products, rather than accounting for physical processes that shape snow accumulation. However, recent work by Manepalli et al. (2019) used a conditional generative adversarial network (cGAN) to emulate VIC-based estimates of SWE developed by Livneh et al. (2015). They formulated this task as an image-to-image translation problem, where the cGAN model translates gridded relationships between the input meteorological fields to the target SWE field without the need for snow-related products. Although the cGAN model is demonstrably powerful, this type of image translation task does not allow time dependency to be incorporated into the model. Namely, it assumes the SWE at time t can be expressed as a function of meteorological variables at the concurrent time t. Under such an architecture, the model cannot capture temporal features from the input predictors, (i.e., the snow accumulation process is ignored), which is vital for time series prediction.

There have also been recent efforts to estimate SWE based on precipitation (P), temperature (T) and other factors that leverage physical causation and a process-based understanding of the system. These new DL models have modeled SWE as an accumulation process by relating SWE to a historic time series of meteorological variables, with the inputs from previous time steps:

$$\mathrm{SWE}_t = f(P_t, P_{t-1}, ..., P_{t-N+1}, T_t, T_{t-1}, ..., T_{t-N+1}) \tag{1}$$

where t denotes the time step and N is the length of the look-back window size. Using the above formula, Meyal et al. (2020) inputted precipitation, temperature, snow depth and SWE from previous days into a long-short-term memory (LSTM) model for SWE prediction at five observational stations. They found that the LSTM model can capture the temporal features of snow accumulation and perform well at the selected stations. Similarly, in Y.-H. Wang et al. (2022) an LSTM model is trained to emulate a gridded SWE dataset, demonstrating the superior ability of LSTM to capture snowpack dynamics over the western US. In these studies, both Manepalli et al. (2019) and Y.-H. Wang et al. (2022) emulated existing SWE products, while Meyal et al. (2020) used observational records and thus did not assume the quality of any existing model or dataset.

Although ML and DL models can achieve satisfying results for historical SWE, models generally struggle with poor performance under extrapolation. Although the LSTM model in Meyal et al. (2020) performed well at the selected observational sites, it was not tested in out-of-sample areas, especially where the statistical properties of SWE accumulation are different from the training sites. This poses a major challenge, particularly if we want to generate a gridded SWE dataset with ML or DL models trained on in-situ observations. Given that in-situ estimates of SWE are generally located in those mountainous areas that are easily accessible and found at mid-elevation, they do not fully represent the areal heterogeneity of SWE at high-elevation or low-elevation that surround the stations (Blöschl, 1999). Therefore, a significant extrapolation problem may arise, particularly when applying the ML or DL models to low-elevation plains or valleys. This issue also makes it difficult to validate or calibrate process-based models, suggesting a need for more observations at both low- and high-elevation. In the case of ML-based models, efforts to address the extrapolation problem include a transformation of the output target for climate emulation or by evaluating model performance using (extreme) out-of-sample scenarios for streamflow projection (S. Duan et al., 2020; Beucler, Pritchard, Rasp, et al., 2021).

In our study, we investigate the viability of DL models for modeling SWE at pointwise locations and as a gridded product. Such datasets would have significant value to both researchers and practitioners, particularly those invested in water resource availability and management. We first build three DL models based on equation (1), only using the meteorological forcings from 581 observational stations in the western United States (WUS). The model behavior and input sensitivity are subsequently analyzed using an explainable artificial intelligence (XAI) method. With these trained DL models in hand, we then tackle the spatial extrapolation problem and generate a gridded SWE product over the Rocky Mountains with 4km grid spacing. This work further sets the stage for a successive effort to leverage our DL model for predicting the response of mountain snowpack to climate change.

The structure of this paper is as follows. Section 2 describes the models employed, the data sources used in our study, and methods for analysis. Section 3 provides a comparative assessment of model performance, including model behavior under cross-validation, and a permutation-based analysis of the DL model to understand which variables are deemed most relevant for SWE prediction. The DL model is then extnded to generate a gridded SWE product, which is described and analyzed in section 4. A discussion of DL model performance in contrast with a process-based model are in section 5, followed by conclusions in section 6.

2 Models, Data and Methods

2.1 Deep Learning Models

Three different DL models applicable to time series problems are investigated and compared, following the general framework depicted in Figure 1. Under this design, the temporal block extracts temporal features from the input data, while the dense layer generates a single-step prediction. The DL models are trained to minimize an objective function (i.e., the loss function), which in this study is chosen to be the mean squared error (MSE). The number of training periods (epochs) is set to 50. The optimization algorithm is Adam with a learning rate of 1e-4 (Kingma & Ba, 2014). Since a gradient-based method is used to optimize the DL model, the converged model will be sensitive to the choice of initial weights. This effect is mitigated by training models 10 times with different initial weights to generate an ensemble of predictions and use the ensemble mean, following X. Wang et al. (2021). The remaining hyperparameters for each model architecture are determined by grid search (more details in Appendix Appendix B). Hyperparameters are not fine-tuned in this study due to the steep computational cost and the minimal benefit awarded by such an approach. All DL models are implemented using PyTorch (Paszke et al., 2019). Specific details on the three DL models, along with our design choices, are as follows.

2.1.1 Long-Short Term Memory (LSTM)

Long-Short Term Memory models (LSTMs) (Hochreiter & Schmidhuber, 1997) are a type of recurrent neural network that has commonly been used in hydrological prediction (Kratzert et al., 2019a; Feng et al., 2020; Lees et al., 2021). LSTMs have demonstrated considerable success for problems of this type, since they are designed to capture temporal dependencies that are common in time-series data.

Details on the mathematical structure of the LSTM are provided in Text S1. The gated design of LSTM enables it to keep and drop information from the previous time steps, which is naturally suited for time series tasks. A detailed figure representing the gates and outputs is depicted in Figure S1. Theoretically, there can be multiple LSTM layers stacked in a single LSTM model. However, the majority of past hydrological application studies adopt a one-layer LSTM model (Kratzert et al., 2019a; Xiang et al., 2020; Feng et al., 2020; Wunsch et al., 2021). In this study, we also utilize a one-layer LSTM model with the number of hidden units (i.e., the dimension of cell state) selected by hyperparameter search.

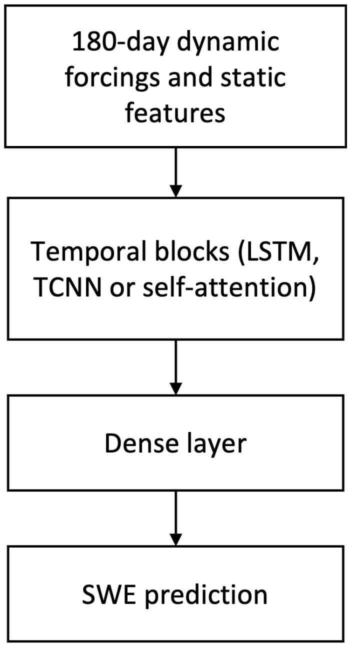

Figure 1. The general framework employed in this study for all ML models.

2.1.2 Temporal Convolutional Neural Network (TCNN)

Historically, convolutional models have been used for image-related tasks because of their ability to extract features with 2- or 3-dimensional convolutional kernels (i.e., weighted inner products that are marched across the input image). Two well-known image-related models built with convolution layers are VGG-16 (Simonyan & Zisserman, 2014) and GoogLeNet (Szegedy et al., 2015). Temporal convolutional neural networks (TCNNs) (Lea et al., 2017), where kernels are instead applied over the time dimension, have also been developed for time-series problems. Bai et al. (2018) tested these models for a variety of standard time series tasks and showed that convolutional models can often outperform LSTMs. TCNNs have also been used in Earth system modelling for predictions of streamflow and the El Niño Southern Oscillation (ENSO) (S. Duan et al., 2020; Yan et al., 2020).

To mimic the inherent time dependencies built into LSTMs, TCNNs use dilated causal convolutions and residual connections (Bai et al., 2018). This architecture is depicted in Figure S2. The causal convolution ensures that outputs at a given time step are only dependent on previous time steps, in contrast to a traditional convolution which could involve future information. The dilated convolution enlarges the receptive field by regularly skipping input time steps; consequently, with stacked deep CNN layers, the receptive field at the final layer can cover the whole input time series. Residual connections are needed along with the stacked layers since the model can be too deep to converge, and residual connections can avoid vanishing or exploding gradients (K. He et al., 2016).

In this study, we use a stacked TCNN model analogous to those employed in Bai et al. (2018) and S. Duan et al. (2020), where each TCNN block consists of two CNN layers and one residual connection (Figure S2). The number of CNN kernels, TCNN blocks and kernel sizes are determined by hyperparameter search.

2.1.3 Self-Attention Model

Vaswani et al. (2017) introduced the Transformer, a self-attention based encoder-decoder model (Bahdanau et al., 2014) for natural language processing (NLP). Since then, many self-attention-based models have been designed and investigated for application to time series problems (Devlin et al., 2018). In Earth system modeling applications, self-attention-based models have been used to predict the ENSO index (Ye et al., 2021) and forecast seasonal precipitation (Civitarese et al., 2021). In recent years, significant effort has been made to optimize the original Transformer architecture and make it more computationally and memory efficient (Tay et al., 2020; Lin et al., 2021). These variants of Transformer models could provide more choices for Earth system applications.

The equations governing the self-attention model are provided in Text S1. In the encoder portion of Transformer, the input vectors are embedded in a dense layer (also called an embedding layer). The self-attention layer takes the embedded inputs and extracts the temporal features, which are then used as input for the decoder (depicted in Figure S3). It can be viewed as a fully connected layer but with dynamical weights representing the pairwise relationships of the input time steps (Lin et al., 2021).

In this study, the encoder of the original Transformer model from Vaswani et al. (2017) is used, featuring a multi-head self-attention mechanism. The number of Transformer encoder layers, number of heads, embedding size and feedforward dimensions are tuned using a hyperparameter search.

2.2 Training Data

Snow Telemetry (SNOTEL) stations provide daily SWE measurements and are used as the prediction target for the ML model. From the 829 stations with available data (including Alaska), we select 581 stations across the WUS with at least one year of complete observations over the training period from 1980 to 1990. Meteorological fields are from the 1/24th-degree ($\sim$4-km) gridMET dataset, including daily precipitation, maximum and minimum temperature, solar radiation, maximum and minimum relative humidity, specific humidity, vapor deficit and wind speed (Abatzoglou, 2013). Since SNOTEL stations do not coincide with gridMET grid points, the data point nearest to each SNOTEL station provides the corresponding forcing.

Static features at each station include latitude, longitude, elevation, diurnal anisotropic heat index (DAH) (Böhner & Antonić, 2009) and topographic solar radiation aspect index (TRASP) (Roberts & Cooper, 1989). DAH and TRASP are used to account for surface solar radiation loading (i.e., shading) (Cristea et al., 2017). DAH is given by

$$\mathrm{DAH} = \cos(\alpha_{\max} - \alpha) \times \arctan(\beta) \tag{2}$$

where $\alpha_{\max}$ is the aspect receiving the maximum amount of solar radiation (for the WUS, we use $\alpha_{\max} = 1.125\pi$, following Böhner and Antonić (2009)), α is the aspect (in radians), and β is the topographic slope (also in radians). DAH ranges between -1 and $+1$, with zero corresponding to flat terrain; for the WUS, DAH is largest on steep southwest-facing slopes that have higher afternoon solar radiation loading and lowest on steep north-facing slopes. TRASP is given by

$$\mathrm{TRASP} = \frac{1}{2}\left[1 - \cos\left(\alpha - \frac{\pi}{6}\right)\right]. \tag{3}$$

TRASP is only a function of topographic aspect and accounts for daily solar radiation loading and ranges between 0 (for the coolest slopes) and +1 (for the warmest slopes). Both TRASP and DAH were calculated using the United States Geological Survey (USGS) Digital Elevation Model (DEM) dataset at 30-meter horizontal resolution. As with gridMET, the nearest grid cell to the SNOTEL station is used as the corresponding input to the DL model.

2.3 Splitting the Data

For purposes of constructing the primary DL models, the data are split into training, validation and testing sets. Several such splittings are used throughout our paper in order to test the robustness of the DL method for capturing snowpack dynamics among different time periods and in different regions. For all splittings, we calculate the mean $\overline{x}$ and standard deviation σ of both the input and output variables in the training period and so normalize the data via

$$X_{\text{normalized}} = \frac{x_i - \bar{x}}{\sigma}. \tag{4}$$

The splittings employed are as follows:

(1) For the temporal train-test split, we use 1980 Oct 1st to 1999 Sep 30th as the training period, 1999 Oct 1st to 2008 Sep 30th as the validation period and 2008 Oct 1st to 2018 Sep 30th as the testing period. All SNOTEL stations are included in this splitting. Since validation is only used to determine hyper-parameters (which are fixed thereafter), this is the only splitting that includes a validation period.

(2a) For the first spatial train-test split, SNOTEL stations are split into eight mountain ranges, including the Pacific Northwest, the Sierra Nevada, the Blue Mountains, Idaho/Western Montana, Northwestern Wyoming, Utah, Colorado, and Arizona/New Mexico. This division follows Serreze et al. (1999) and M. He et al. (2011b), where it was shown that these eight mountain ranges exhibited distinct snow dynamics. This splitting includes eight experiments, in each case using seven mountain ranges for training and one for testing.

(2b) For the second spatial train-test split, all SNOTEL stations are randomly split into eight subsets or folds. Each time this splitting is performed, seven folds are used for training and the rest for testing. Unlike the spatial splitting in (2a), this spatial splitting still allows the model to comprehensively learn snow dynamics from throughout the western US.

2.4 Gridded SWE Reference Products

Spatiotemporally complete observations of SWE in mountain areas remain elusive, requiring us to instead employ reconstructions that meld models and observations. Of course, such products inevitably inherit biases from incomplete observations and uncertainties in the model design, particularly in regions where observations are sparse. To better quantify these structural uncertainties, two model products are employed in this study. These two products were chosen because they are modern, high-quality data products that are widely used in the snow modeling community. Other such SWE products can be found in McCrary et al. (2017) and Snauffer et al. (2018).

The primary product employed is the daily 4km gridded SWE data from Zeng et al. (2018) (hereafter referred to as the University of Arizona or UA dataset), which uses PRISM precipitation and temperature data and assimilates SNOTEL observations. In the UA product, rainfall and snowfall are partitioned using daily 2m air temperature thresholds derived from station observations. When interpolating point measurements to a grid, the ratio of SWE observations is used instead of the absolute SWE measurements as net snowfall was found to be overestimated. Further details on the methodology employed and corresponding analysis can be found in Broxton et al. (2016) and Zeng et al. (2018).

The second product adopted in this study is an independent SWE dataset developed at the University of California, Los Angeles, (referred to as the UCLA dataset). The UCLA dataset takes three Landsat sensors as input, along with meteorological forcings, topographical features and landcover data. The snow estimates are then updated with MODIS remote sensing estimates of snow cover too. Within a Bayesian framework, this dataset provides ensemble statistics of SWE estimates (e.g., mean, standard deviation, median) (Y. Fang et al., 2022). Details about the processing algorithm can be found at Margulis et al. (2019). The horizontal resolution of the UCLA product is 16 arc seconds, which varies from 350m to 500m. For the purposes of this study we use the ensemble mean SWE estimate, which is regridded to the same 4km resolution grid as the UA dataset.

2.5 The SNOW-17 Model

One issue with the use of gridded products is that they do not provide SWE data at the precise SNOTEL station locations. Interpolating meteorological and SWE data from gridded data points to SNOTEL stations can introduce potentially significant errors, particularly in regions of complex topography. Consequently, we further compare our DL-based SWE estimates to those from the process-based SNOW-17 snow accumulation and ablation model. SNOW-17 uses an air temperature index to determine energy exchanges at the snow-air interface and enforces principles of water and energy balance to estimate SWE and runoff. We refer readers to E. A. Anderson (1976) and E. Anderson (2006) for the detailed processes and equations used in the SNOW-17 model.

Training data from selected SNOTEL stations and gridMET are used to calibrate the SNOW-17 model (i.e., 1980 Oct 1st to 1999 Sep 30th). The model is then used to generate SWE estimates over the testing period (2008 Oct 1st to 2018 Sep 30th) for comparison. Candidate tuning parameters are determined based on previous studies on model sensitivities (e.g., E. A. Anderson, 1973; M. He et al., 2011a; Raleigh & Lundquist, 2012) and listed in the Appendix. The shuffled complex evolution approach (SCE-UA) is used to optimize the parameters, with details in Q. Duan et al. (1993, 1994).

2.6 Performance Metrics

Model performance is quantified using the Nash-Sutcliffe model efficiency coefficient (NSE), a widely used metric for hydrological model evaluation (Nash & Sutcliffe, 1970). It is defined via

$$\mathrm{NSE}(O_t, P_t) = 1 - \frac{\sum (O_t - P_t)^2}{\sum (\overline{O}_t - O_t)^2}, \tag{5}$$

where O and P denote observations and predictions, respectively. Index t denotes the time and $\overline{O}_t$ is the observation mean. NSE is in the range $(-\infty, 1]$, with larger values indicating better performance and a score of 1 indicating a perfect match between model and observations. Note that the NSE score is not symmetric, i.e., $\mathrm{NSE}(A, B) \neq \mathrm{NSE}(B, A)$; in this study the first NSE argument consistently refers to the reference product.

In this study, we employ NSE in two ways. First, the NSE of absolute SWE is calculated as

$$\mathrm{NSE}_{\mathrm{absolute}} = \mathrm{NSE}(\text{SWE-REF}, \text{SWE-DL}), \tag{6}$$

where SWE-REF denotes the SWE from the reference dataset and SWE-DL denotes the SWE prediction from the deep learning model. Second, the NSE of the SWE fraction is given by

$$\begin{aligned} \mathrm{NSE}_{\mathrm{fraction}} &= \mathrm{NSE}\left(\frac{\text{SWE-REF}}{\max(\text{SWE-REF})}, \frac{\text{SWE-DL}}{\max(\text{SWE-DL})}\right), & (7) \\ &= \mathrm{NSE}\left(\text{SWE-REF}, \frac{\text{SWE-DL}}{\max(\text{SWE-DL})} \times \max(\text{SWE-REF})\right) & (8) \end{aligned}$$

where max(SWE-REF) represents the historical maximum SWE from the reference dataset, and max(SWE-DL) denotes the historical maximum SWE from the DL models. Equations 7 and 8 are equivalent since the NSE value is unaffected when the predictions and observations are multiplied or divided by the same constant. As opposed to the NSE of absolute SWE, the NSE of SWE fraction emphasizes the temporal features and de-emphasizes errors in magnitude.

Model performance is further quantified using root mean squared error (RMSE) and mean absolute error (MAE),

$$\begin{aligned} \mathrm{RMSE}(O_t, P_t) = &\quad \sqrt{\frac{\sum (O_t - P_t)^2}{n_{\text{samples}}}}, & (9) \\ \mathrm{MAE}(O_t, P_t) = &\quad \frac{1}{n_{\text{samples}}} \sum |O_t - P_t|. & (10) \end{aligned}$$

where n_{samples} is the number of evaluated samples. RMSE and MAE are in the range $[0, +\infty)$ with lower values indicating a closer match, and a score of 0 indicating a perfect match between the model and observations.

2.7 Feature Permutation

Although DL models generate accurate predictions, they are frequently referred to as 'black box' models since it is often unclear why and how the model produces its results. Recent advances in explainable AI (XAI) have enabled better interpretation of DL model results, especially in Earth system modeling (McGovern et al., 2019; Gagne II et al., 2019; Barnes et al., 2020; Toms et al., 2020). Such techniques are further useful for building credibility in DL models by demonstrating that they are mimicking physical understanding and principles.

Permutation-based XAI methods are commonly used to quantify the relative importance of input variables in the DL models (Breiman, 1996). The permutation method evaluates the DL model by first obtaining a baseline performance score. Then each feature is permuted to generate a shuffled dataset, and a new performance score is calculated. The change in the performance score represents the importance of a given feature. A greater decrease in model skill corresponds to higher feature importance. This approach follows previous work addressing model interpretation (Gagne II et al., 2019). However, care should be taken in the interpretation of these results, as the quantified performance is potentially confounded by correlation among input features. For example, a model that uses both mean and maximum daily temperature as input may see minimal performance loss from the removal of either of these features while the removal of both would be significant. Efforts to address correlation issues include the use of multi-pass permutation, as discussed in a review by McGovern et al. (2019).

In this study, we permute both the training and the testing set and train a reduced model. By permuting the training set and retraining the model, the permuted variable is blocked and the reduced model only receives the information from the remaining non-permuted variables. The importance of the permuted variable will be quantified by examining the ratio of the new NSE value against the baseline score. The permutation is performed separately for dynamic inputs and static variables. For dynamic inputs, the time series from each grid point is used for re-sampling so that the statistical properties of these variables are preserved (i.e., only the time steps are shuffled). For static features, the permutation is performed among all stations.

2.8 Switching the Model Target for Spatial Extrapolation

DL models generally yield accurate predictions when interpolating between unseen samples in the training set range, but can struggle when extrapolating beyond this range. With that said, Balestriero et al. (2021) showed that in high-dimensional data with the training range defined by the convex hull of the training set (i.e., the minimal convex polygon that encompasses all the training points), samples almost always fall in the extrapolation regime. This is particularly true for time series problems, where the dimension is the product of input time window size and a number of input features. For the problem investigated here, extrapolation is most obvious when the SNOTEL-trained model is applied to the whole Rocky Mountains, since these stations are largely found only in high-elevation regions. To mitigate some of the effects of extrapolation, an alternate model target is considered in this work.

To alleviate problems with extrapolation, a second set of DL models are trained with SWE fraction as a target, which is defined as the SWE normalized by the historical maximum SWE at each station. In this case, the model output is generally less than 1, though exceedance of maximum historical SWE is allowed and would produce values greater than 1. Consequently, for the SWE fraction, model normalization is not needed (i.e., equation 4). When extrapolating to a low-elevation area, the DL estimation is compared against the reference dataset. The absolute SWE can be recovered by multiplying the model result by the historical maximum SWE. In this case, the NSE of SWE fraction (equation 8) can be written as

$$\begin{aligned} \mathrm{NSE}_{\mathrm{fraction}} &= \mathrm{NSE}(\text{SWE-REF}, \text{SWE-DL-FRAC} \times \max(\text{SWE-REF})) && (11) \\ &= \mathrm{NSE}\left(\frac{\text{SWE-REF}}{\max \text{SWE-REF}}, \text{SWE-DL-FRAC}\right), && (12) \end{aligned}$$

where SWE-DL-FRAC denotes the DL model predicted SWE fraction.

3 Model Performance at SNOTEL stations

3.1 Computational Performance

The training time for each DL model on a single RTX 2080TI GPU is 5 hours for the LSTM model, 10 hours for the TCNN model, and 26 hours for the Attention model. Although the training time varies among different DL models, all DL models only need to be trained once, and inference time is much shorter. Inference time at SNOTEL stations is on the order of a few minutes. For our application over the Rocky Mountains, it takes approximately 35 minutes to generate a 10-year prediction of SWE on the 168 $\times$ 108 grid cells covering the Rocky Mountains with parallel execution on a single RTX 2080TI GPU for either the LSTM and TCNN. The Attention model takes longer than the other DL models to generate a prediction – approximately 2.5 hours without parallel execution – and is limited by the GPU memory. Faster performance is anticipated using GPUs with larger memory. Given the much faster inference time, the training time is not used for model selection or intercomparison.

3.2 Temporal Prediction

Following DL model training, DL model performance is evaluated over the testing period (2008-10-01 to 2018-09-30) at all SNOTEL stations. Since summertime SWE is zero at most SNOTEL stations, daily information for both the entire testing period and only those days in the testing period when observed SWE is nonzero are examined separately. As discussed in section 2.1, the mean from a 10-member ensemble SWE prediction for each DL model and SNOTEL station is used for analysis. Table 1 shows the median NSE, MAE and MSE across all SNOTEL stations. Among the DL models, the LSTM has the highest median NSE value, followed by the TCNN and then the Attention model. The distribution of NSE values is further shown in Figure 2. The LSTM cumulative distribution function (CDF) lies above the TCNN and Attention CDF, indicating that the LSTM produces more station SWE predictions with higher NSE values, compared with the TCNN and Attention models. When only comparing the TCNN and Attention models, the TCNN has a slightly higher median performance, although the Attention model shows better overall performance. An example of SWE prediction for a single SNOTEL station is given in Figure S5.

Prediction accuracy is also computed for the UA and the UCLA dataset over the same testing period using the nearest grid point to the SNOTEL station (further dis-

Table 1. Tabulated model performance for prediction of SWE at SNOTEL stations. The top table shows performance scores on dates when observed SWE is greater than zero, while the bottom table shows the whole evaluation period (from 2008 to 2018). The best scores for each metric are shown in bold font.

Nonzero SWE	Median NSE	Median MAE (mm)	Median RMSE (mm)
LSTM	**0.823**	**42.97**	**63.14**
TCNN	0.792	52.58	73.14
Attention	0.779	52.03	72.06
UA	0.755	49.77	73.02
UCLA	0.482	77.31	104.51
SNOW-17	0.488	80.34	102.39
Whole period	**Median NSE**	**Median MAE (mm)**	**Median RMSE (mm)**
LSTM	**0.901**	**25.76**	**49.03**
TCNN	0.879	31.07	55.76
Attention	0.871	29.93	54.08
UA	0.861	26.69	53.61
UCLA	0.708	47.16	81.19
SNOW-17	0.722	44.69	77.45

cussion of the potential impact of this choice on accuracy is provided in section 5). The median NSE values over the whole testing period are 0.861 and 0.482 for the UA and UCLA dataset, respectively. For those days when observed SWE is nonzero, the NSE score is 0.755 for the UA dataset and 0.708 for the UCLA dataset. Under this metric, all DL models achieve better results than these two datasets. Moreover, when comparing the distribution of NSE values in Figure 2, we see that the DL models produce more stations with higher NSE values. As for MAE and MSE metrics, the UA dataset achieves a lower median MAE than the TCNN and Attention for both the whole testing period and the nonzero-SWE period. Examining Figure 2, the UA dataset also attains the lowest MSE and MAE across 20% to 30% of SNOTEL stations.

Finally, prediction accuracy from the DL models is compared with the SNOW-17 model. Under all three metrics, all DL models attain higher NSE scores and lower MAE and MSE values than SNOW-17. Since we used the same training period to calibrate the SNOW-17 model, missing values may affect model parameters at some locations. In fact, the median NSE score for the SNOW-17 model increases to 0.778 when only those stations with less than one year of missing values are taken into consideration. Nonetheless, this score is still lower than the DL models. Possible reasons are discussed further in Section 5.

To better understand the drivers of model performance, we examine relationships between NSE, elevation, and maximum SWE. Figure 3 shows the NSE distribution with respect to elevation. In general, all DL models and process-based model/datasets perform better at higher elevation. For the DL models, performance improvements are also positively correlated with maximum SWE (as shown in Figure 4). This relationship is not obvious among the process-based models, and seems to only hold for maximum SWE below 1500mm (1000mm for UCLA). However, sharp variations in performance can be partially attributed to low sample size for high maximum SWE. This result appears to be a common theme in our study: at higher elevations and in regions of deeper maximum SWE, snowpack is easier to predict. We hypothesize that this is because there is less uncertainty in rain-snow partitioning at higher elevation, and because the processes

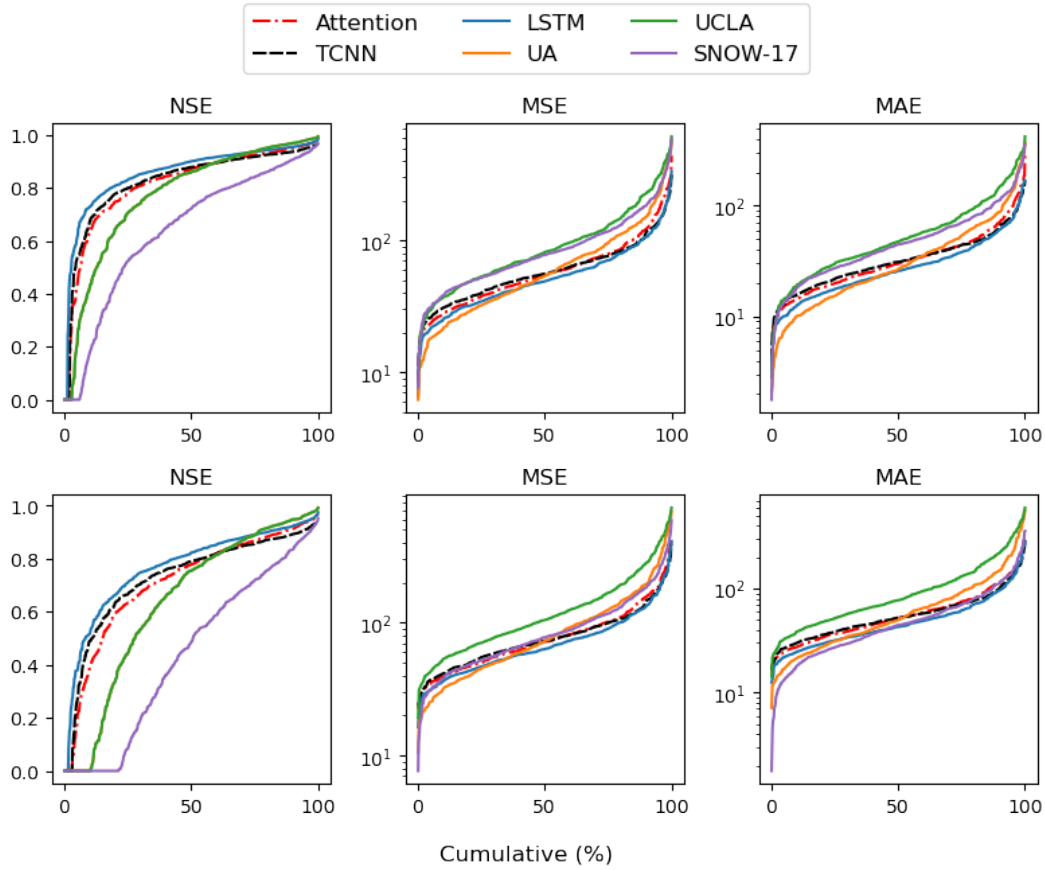

Figure 2. Cumulative distribution functions across three model performance metrics for predicting SWE across SNOTEL stations in the western US. NSE values are truncated at 0. The upper row shows model performance when only those SWE observations larger than zero are used. The lower row shows model performance for all zero and nonzero SWE observations. Higher scores in the first column, and lower scores in the second and third columns, are indicative of higher performance.

that occur at the interface between land surface and snowpack are strongly nonlinear, but become less important when a deep snowpack is present.

To summarize, for western US SNOTEL stations, DL models can generally produce comparable results with the selected process-based model and datasets, and share a similar relationship between performance and elevation. Among all models, the LSTM model provides the most accurate predictions across the three DL models assessed. Although the three DL models exhibit differences in performance, the spatial distribution of their performance tends to be similar – that is, stations with lower (higher) NSE values in one model tend to have lower (higher) NSE values in other models. Across DL models, the Pearson correlation of NSE is 0.856 and 0.660 for TCNN and Attention versus LSTM (Figure S6). As shown in Figure 5, all DL models exhibit relatively poor performance (i.e., negative NSE values) in Western Washington, Northern Nevada, Southern and Northwestern Oregon, and Northern Montana. In general, these stations tend to have a lower maximum SWE than other stations, which is consistent with our earlier attribution of model performance (Figure 4). These also tend to be regions where the UCLA product performs poorest, while the SNOW-17 and UA datasets are more variable. Importantly, these results should not be taken as being indicative of the UA product being higher quality than the UCLA product. While the UA dataset assimilates SNOTEL observations, the UCLA dataset relies almost exclusively on remote sensing, as outlined in Section 2.4. Consequently, the relatively strong performance of the UA dataset is unsurprising when evaluated against SNOTEL observations. A more thorough comparison between the UA and UCLA datasets (and, more generally, observational spread) would require an independent data source (e.g., the Airborne Snow Observatory), which lies beyond the scope of this study.

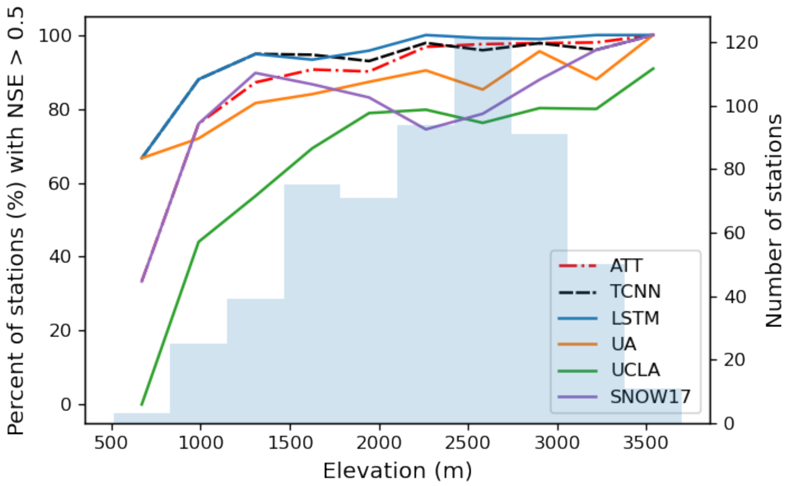

Figure 3. NSE value distribution with respect to elevation. The left y-axis denotes the fraction of stations with NSE values higher than 0.5 for each elevation bin. The right y-axis shows the number of SNOTEL stations in each elevation bin.

3.3 Spatial Cross-Validation

The ability of the model to transfer its understanding of physical processes from one region to another is now assessed, as we build towards the development of a gridded SWE product. Hereafter, our study will focus exclusively on the LSTM because of its superior performance compared with other DL models and its strong correlation with those models across stations. Among 581 SNOTEL stations, 530 are located inside the mountain range boundaries in Serreze et al. (1999) and M. He et al. (2011b). The two spatial splittings employed here are described in section 2.3, and are referred to as 'mountain cross-validation' for splitting (2a) and '8-fold cross-validation' for splitting (2b). The 'time-split' experiment which was analyzed in section 3.2 is used as a reference. Results from this experiment are given in Figure 6 and Table 2. Overall, the 'time-split' LSTM yields the best prediction accuracy, with a median NSE score of 0.899, followed by the '8-fold cross-validation' and 'mountain cross-validation' LSTMs, with the NSE scores of 0.888 and 0.844, respectively. Compared with our full model that trained with 581 SNOTEL stations, we do see a tendency of better performance with more training stations. This suggests the benefit of a large and diverse training set, which was also argued in K. Fang et al. (2022).

Grouped by mountain ranges, Idaho/Western Montana and northwestern Wyoming areas exhibit stronger performance, while the Cascades produces the lowest median NSE score, which is especially pronounced for the 'mountain cross-validation.' This suggests there are unique snow dynamics in the Cascades that other mountain ranges appear unable to capture. In addition, predictability is limited in Arizona/New Mexico (AZ/NW),

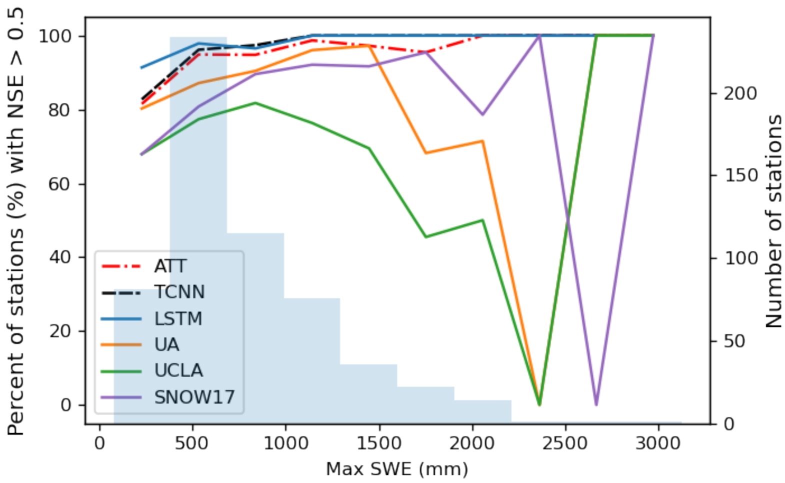

Figure 4. NSE value distribution with respect to maximum SWE measurement. The left y-axis denotes the fraction of stations with NSE values higher than 0.5 for each SWE bin. The right y-axis shows the number of SNOTEL stations in each SWE bin.

Table 2. Median NSE values for SNOTEL stations in major mountain ranges. Northwestern Wyoming is abbreviated to NW Wyoming and Arizona/New Mexico to AZ/NW. Numbers in parenthesis denote the number of SNOTEL stations in each mountain range.

	Cascades (78)	Sierra Nevada (24)	Blue Mountains (26)	Idaho/Western Montana (95)	
Time-split	0.853	0.878	0.862	0.923	
Cross-validation	0.812	0.845	0.879	0.914	
Mountain-based	0.741	0.846	0.852	0.894	
	NW Wyoming (110)	Utah (74)	Colorado (109)	AZ/NW (19)	Overall (530)
Time-split	0.921	0.911	0.907	0.856	0.899
Cross-validation	0.901	0.891	0.892	0.792	0.888
Mountain-based	0.877	0.867	0.833	0.758	0.844

the southernmost of our selected mountain ranges. These results are not surprising given the distinct topographical features of these regions: the elevation is much lower in the Cascades compared with other mountain ranges (shown in Figure 6), while the AZ/NW mountains experience relatively warm temperatures and lower maximum SWE. Indeed, when the AZ/NW mountains are used for testing, elevation and latitude are completely out of the training range. This is an obvious example of model extrapolation, the likely explanation for this range's relatively poor performance, and suggests a need for more observational data from a variety of snow regimes. Nonetheless, the DL model performance in this test exceeds NSE scores derived from SNOW-17 and the UCLA product.

3.4 Permutation-Based Analysis

The variables that are most important for the LSTM model are now studied using a permutation-based analysis. As described in section 2.7, a new set of LSTM models are trained and tested with the shuffled datasets. The importance of each input variable is quantified by comparing the ratio of the permuted LSTM model whole period NSE prediction to the LSTM model baseline NSE prediction, which is trained and tested with the non-permuted dataset. The permuted LSTM models are also trained 10 times to build an ensemble of predictions. To quantify model uncertainty, we use bootstrap sampling

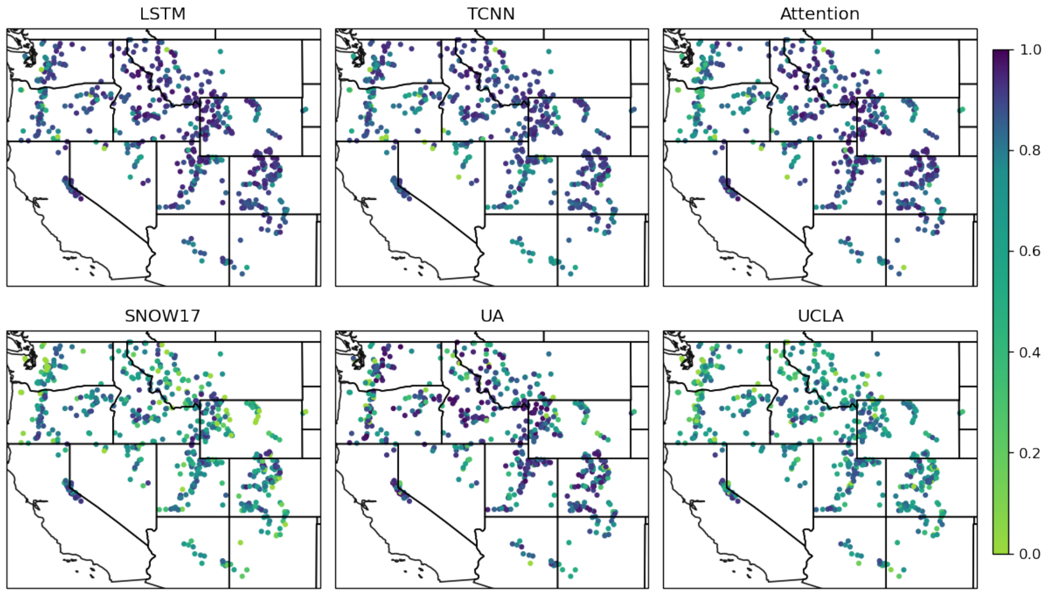

Figure 5. SWE prediction performance from the DL models, the SNOW-17 model, and the UA and UCLA datasets. Dots represent individual SNOTEL stations, with the color of the dot representing the NSE value (truncated at 0).

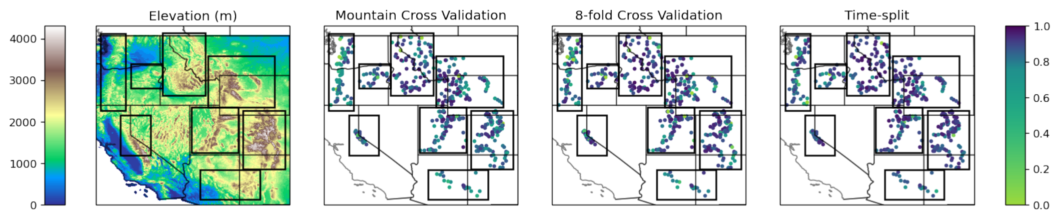

Figure 6. SWE prediction performance from the LSTM with spatial cross-validation. Elevation is shown on the left as a reference. Dots represent individual SNOTEL stations, with the color of the dot representing the NSE value (truncated at 0). Mountain-based cross-validation is depicted on the left, random 8-fold cross-validation in the middle, and the time-split result is shown on the right for reference. Black boxes represent the mountain region boundaries.

to provide results with a 90% confidence interval. In Figure 7, the green bars represent the performance decline for each static variable, the blue bars represent the performance decline from individual meteorological variables, and the orange bar represents the combined effect of all static variables.

The input variable with the most influence on model performance was precipitation, followed by elevation, while the rest of the input variables had comparable influence. This result agrees with the intuition that precipitation provides water mass to build snowpack, and precipitation type is determined by temperature (and humidity), which is shaped by elevation via the lapse rate (Jennings et al., 2018). Although this result seems like common sense and may be affected by the collinearity between input variables, it helps build trust in the LSTM model and provides evidence that it follows basic physical principles. The combined effect of static features is also demonstrated to be critical, as the LSTM model accuracy would drop approximately 7% without their inclusion, which is more than half the influence of precipitation. Clearly, these features are use-

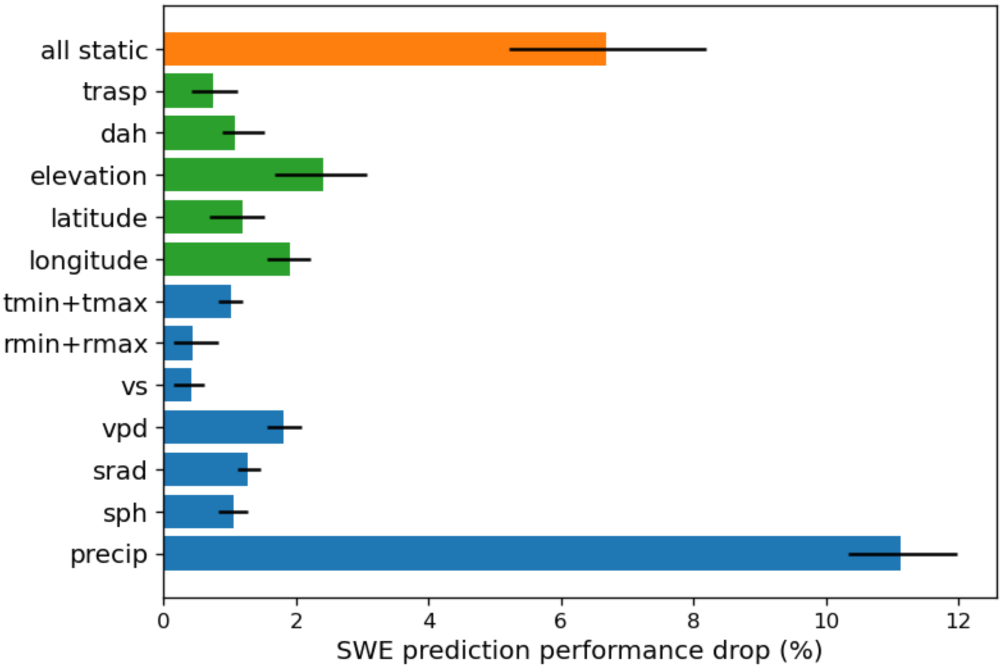

Figure 7. SWE prediction performance drop (%) quantified using the NSE values among the permuted estimates. Error bars represent the 90% confidence interval from bootstrap sampling. Precipitation is abbreviated to 'precip', 'sph' stands for specific humidity, 'srad' for solar radiation, 'vpd' for vapor deficit, and 'vs for wind speed. 'Tmin', 'tmax', 'rmin', and 'rmax' refer to minimum and maximum temperature and relative humidity.

ful for modulating snowpack dynamics at each SNOTEL station and in out-of-sample locations during extrapolation. The utility of static variables was also reported in Kratzert et al. (2019b), where LSTM models were used to predict streamflow.

Among all static variables, there are three categories: location (latitude and longitude), aspect and slope (DAH and TRASP), and elevation. We combined each static variable into these categories during the permutation process to compare their relative importance. Relative to the baseline LSTM model with a median NSE of 0.901, the LSTM model that did not include location information had the highest median NSE score (0.878), followed by aspect and slope (0.874), and elevation (0.870). Despite being rather modest drops, this result again emphasizes that elevation information is the most important in SWE prediction since it can determine the temperature and rain-snow partitioning of precipitation. Although the local temperature is also affected by latitude through differences in solar loading, the LSTM model benefited more from information related to aspect and slope, which have more localized effects on temperature.

Because many of these variables are correlated, care should be taken in attributing the relative importance of variables other than precipitation under the permutation test. This is especially true for temperature, since vapor deficit is a function of temperature and relative humidity and consequently, temperature can be inferred from vapor deficit and relative humidity even if we permute temperature. To better compare their relative influence on SWE predictability, several reduced-order LSTM models are trained. In each reduced-order model, precipitation and one of the other meteorological variables are used, and the remaining variables are permuted. The baseline for comparison was an LSTM model with only precipitation and the reference was a model with the full set

Table 3. First quantile, median and third quantile whole period NSE from several reduced-order LSTMs for predicting total SWE. Precipitation is abbreviated to Precip.

	First Quantile NSE	Median NSE	Third Quantile NSE
Precipitation Only	0.149	0.380	0.576
Precip+Wind Speed	0.432	0.620	0.740
Precip+Specific Humidity	0.686	0.826	0.896
Precip+Solar Radiation	0.709	0.836	0.905
Precip+Vapor Deficit	0.740	0.843	0.894
Precip+Temperature	0.793	0.874	0.924
Precip+Temperature+Relative Humidity	0.784	0.875	0.919
Precip+Temperature+Vapor Deficit	0.818	0.881	0.927
Precip+Temperature+Specific Humidity	0.799	0.882	0.927
Precip+Temperature+Wind Speed	0.803	0.884	0.926
Precip+Temperature+Solar Radiation	0.813	0.886	0.931
Full Model	0.831	0.901	0.938

of meteorological variables. The model using precipitation plus relative humidity was not included in this analysis because it did not converge to a reasonably performant model. As shown in Table 3, among the reduced-order models, precipitation and wind speed give the lowest NSE value, although even this combination does improve skill tremendously compared with the baseline model. The median NSE scores across the rest of the reduced-order models are all above 0.8, and the combination of temperature and precipitation produces the closest performance to the reference model. This indicates that vapor pressure deficit, solar radiation, and specific humidity contain influential information for SWE prediction, while temperature is the most critical variable for model skill besides precipitation.

To determine the best third variable in the model, five additional models were trained: each model consists of precipitation, temperature, and one other variable. The results are shown in Table 3. The model with precipitation, temperature and relative humidity attains the lowest NSE value and is very close to the model with only precipitation and temperature, which is consistent with previously observed anomalous low performance with precipitation and relative humidity. The inclusion of vapor deficit, specific humidity and wind speed all increase the model performance and yield similar NSE scores, probably because these variables cannot be inferred from precipitation and temperature. The model with precipitation, temperature and solar radiation obtained the highest median NSE value of 0.886, or 98% of the full model performance. Clearly, with far fewer input variables, the model with precipitation, temperature and solar radiation was capable of capturing the temporal features necessary for SWE prediction. This again highlights the important roles that these three variables have in affecting the water cycle (S. Duan et al., 2020). Additionally, this result suggests that good estimates of snowpack can be obtained from datasets providing these quantities in high quality, such as CAMELS (Addor et al., 2017).

One additional model was trained to capture some of the diurnal cycle of temperature through inclusion of both minimum and maximum temperature (as opposed to daily average temperature). The improvement in median NSE was only 0.002, suggesting minimal value to the inclusion of both variables (more information in Table S1 and Text S2).

4 Spatial Extrapolation of DL Models to the Rocky Mountains

Our earlier analysis indicates that DL models are capable of predicting daily SWE at individual SNOTEL stations and can even achieve satisfactory performance when extrapolating to stations out of the training set. A gridded SWE product similar to the UA and UCLA datasets is now developed by applying these models out-of-sample across the Rocky Mountains at 4km grid spacing (Figure 8). It is shown that the resulting product is reasonable, even when there are out-of-sample differences in the statistical properties of the DL models' input and output variables. The use of these models outside of their training range is a common problem referred to within the machine learning community as concept drift or extrapolation (Tsymbal, 2004). In this case, extrapolation is expected to be common since, in addition to other differences, many grid points have elevations lower or higher than the lowest or highest SNOTEL station (this was also hypothesized to have impacted model performance over the Cascades and New Mexico in Table 2).

The 4km grid used in this application is inherited from the gridMET forcing data (section 2.2). A similar approach could also be used to produce an even higher resolution product (e.g., one matching the 800m Parameter-elevation Regressions on Independent Slopes Model (PRISM) product). The simulation period is 2008-10-01 to 2018-09-30, the same as the SNOTEL testing set. For better comparison across different spatial resolutions, the gridMET data is also regridded to the UA and UCLA grid points using the nearest neighbor method. When applied over the Rocky Mountains, the gridMET forcing variables are normalized with the mean and standard deviation from the training SNOTEL stations (equation 4). The DL model SWE prediction is then transformed back to its original units with the same equation. The top row in Figure 9 and 10 shows the NSE values obtained from the DL-generated dataset (10 ensemble member mean) when using the UA dataset and UCLA dataset as reference, respectively, following equation 6.

The DL model estimates largely agree with the process-based estimates in high-elevation areas, while performance is relatively poor in low-elevation areas. Given that much of the domain is covered by low-elevation areas, it is useful to investigate the reasons for this poor performance and develop models which can mitigate these errors. One obvious driver of poor performance is the sensitivity of NSE to differences that are relatively small in absolute magnitude, when maximum SWE is already small. This is illustrated by comparing the UA and UCLA datasets (middle and right figures in Figure 8). Negative NSE values abound in low-elevation areas (e.g., the northwestern Rockies, 36°N-37°N and 109°W-108°W), suggesting significant disagreement between these two products in this region. This difference also appears in Figure 3, where both DL models and the UA and UCLA datasets exhibit poor performance at lower elevations and when SWE amounts are low – indeed, the ground truth in these regions is poorly constrained given a dearth of relevant measurements. The discrepancy between UA and UCLA in this region is likely exacerbated by the employ of different algorithms and data sources: UA is not informed by remote sensing estimates, but is informed by SNOTEL stations, while the opposite is true for the UCLA product. These factors make it difficult to quantify how much error may be attributed to the out-of-sample application of the DL model.

To better understand the reasons for poor performance in low-elevation areas, errors in this region are decomposed into errors in magnitude estimation (i.e., too little or too much SWE) and errors in temporal dependency (i.e., too slow/rapid accumulation/melt). To mitigate issues related to magnitude estimation, model performance is assessed using the fraction of maximum SWE (i.e., $\mathrm{SWE}/\max \mathrm{SWE}$), where the maximum is with respect to the historical/training period. That is, the NSE from snow fraction is calculated via equation 7. By using the SWE fraction, differences in SWE magnitude between the reference datasets and DL model are mitigated and so the evaluation emphasizes the temporal character of the SWE (e.g., the timing of accumulation

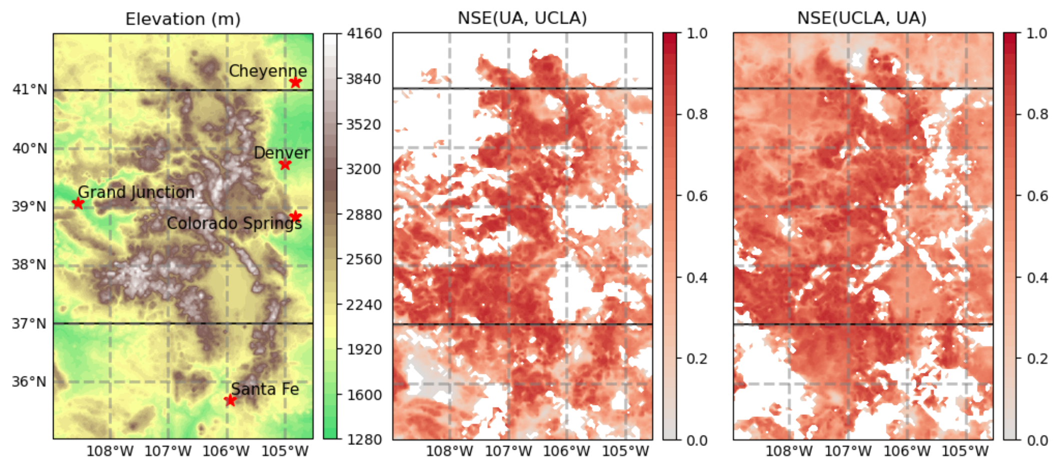

Figure 8. Elevation (left) along with NSE scores between reference datasets with UA as reference (middle) and UCLA as reference (right) over the Rocky mountain area.

and melt). The middle row in Figure 9 and 10 shows the assessment with SWE fraction. Under this metric, the DL model appears significantly better when evaluated against the UA dataset in Figure 9, with higher NSE values almost everywhere and a larger portion of positive NSEs. This difference indicates that while the DL models can capture the temporal dependence of SWE, magnitude biases can be relatively large over low-elevation areas. However, when the UCLA dataset is used as a reference, the fractional SWE metric does not always lead to improvements in the performance of the DL models (Figure 10): in fact, only the TCNN model produces more grid points with positive NSE values, indicating some temporal feature mismatch between the LSTM and Attention models and the UCLA dataset that cannot be mitigated under this metric. Despite the overall decrease in the fraction of positive NSE scores for the LSTM and Attention models, there are indeed improvements over the mountain range to the northeast of Santa Fe (36N-37N, 106W-105W). This pattern of improvement is consistent across all the DL models and independent of the reference dataset, as similar patterns are also observed in Figure 9.

Given the improvement in model performance when using SWE fraction, a new set of DL models is trained on SNOTEL data to predict SWE fraction (section 2.8), rather than the SWE itself (with predictions hereafter referred to as SWE-DL-FRAC). NSE values are then computed using equation 12. It should be noted that equation 8 and 12 are not equivalent and they represent two different comparisons. Equation 8 uses the original DL models, which predicts SWE magnitude normalized by the historical maximum prediction; whereas the maximum SWE in equation 12 is from reference datasets and the DL models are predicting SWE fractions. The bottom row in Figure 9 and 10 show the NSE results when predicting SWE fraction directly from the DL models. When compared to the original DL models which predict SWE magnitude, the DL models trained to predict SWE fraction exhibits a clear and significant improvement almost everywhere in the domain, but particularly in low-elevation regions. This result shows that normalization by maximum SWE is effective for all the DL models with both reference datasets. Among all the DL models, the LSTM-based model again provides the best overall SWE prediction, determined by the largest fraction of positive NSE values. Of course, to transform the fraction of maximum SWE back to an absolute SWE value, the historical maximum SWE is needed within each grid cell. Since SNOTEL observations are unevenly distributed throughout the Rocky Mountains, we must rely on an alternative estimate of maximum SWE at each grid point; in this case, we use the historical maximum SWE

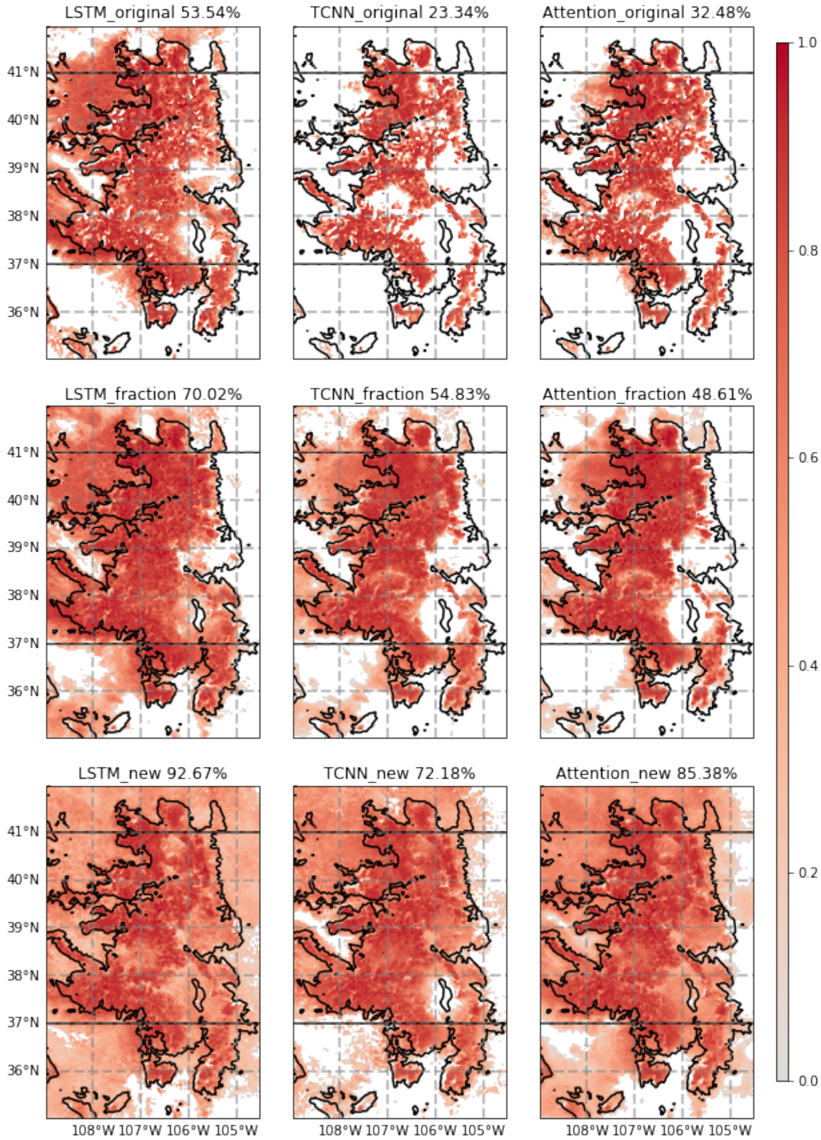

Figure 9. NSE values for DL model extrapolation estimates over the Rocky Mountains with the UA dataset as reference. The top row shows the NSE score of the original DL model SWE predictions. The middle row is the SWE fraction evaluation from the original models, computed via equation (7). The bottom row represents the new set of DL models that predict SWE fraction, computed via equation (12). NSE values below 0 are masked in all figure subpanels. The black line is the 2300-meter contour. The percentage value given in the title is the fraction of grid points with positive NSE values.

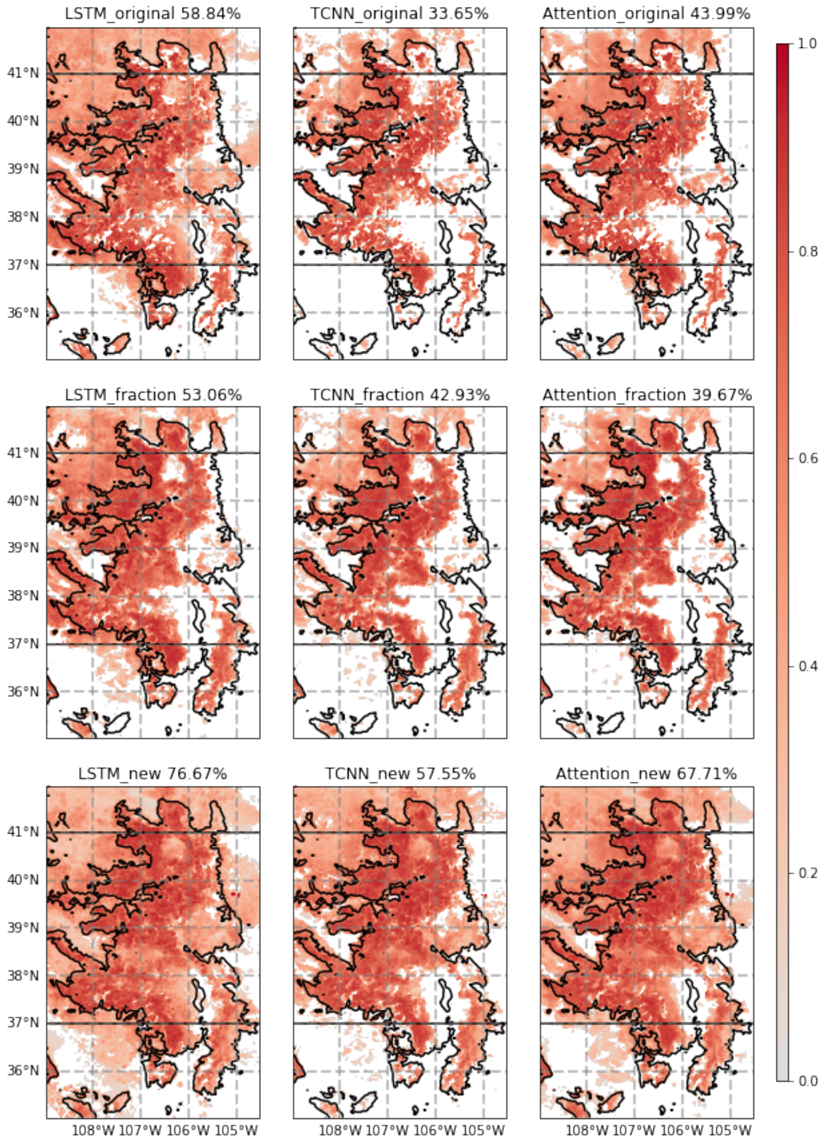

Figure 10. Same as Figure 9, but with the UCLA dataset as reference.

values from the reference dataset, either the UA or UCLA dataset, at each grid point over the training period to estimate maximum SWE. An example of annual maximum SWE prediction from this new LSTM model is shown in Figure S7.

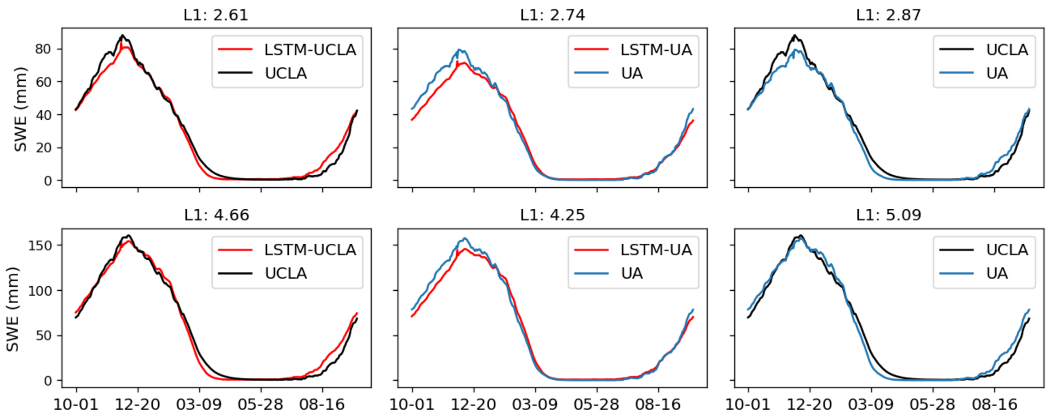

Figure 11. Area-averaged SWE climatology over the Rocky Mountain area. The first row depicts results from the whole region, while the second row depicts only the area above 2300m. 'UCLA' and 'UA' stand for the results from the reference datasets. SWE fraction from LSTM is transformed back to SWE depth and the results are denoted with the corresponding reference datasets. The title for each subfigure denotes the L1 norm between the two climatology estimations.

LSTM performance is further analyzed by examining the annual SWE climatology over the Rocky mountains. The SWE estimation in millimeters, which is derived from the DL-generated SWE fraction and reference datasets as in equation 12, is averaged over the Rocky Mountain area and compared against the reference dataset. In Figure 11, 'LSTM-UCLA' is derived from the LSTM estimation and historical maximum SWE from the UCLA dataset, whereas 'LSTM-UA' uses the UA dataset to derive the historical maximum SWE. In general, the LSTM model matches the reference dataset, with correlation coefficients exceeding 0.99. The LSTM model does, however, have a tendency to underestimate the snowpack peaks, but this bias decreases in higher-elevation areas. The magnitude difference between these two reference datasets is also worth noting. Averaged over the whole area, a higher peak SWE is observed in the UCLA dataset, which again tends to diminish with elevation. This suggests that both process-based and DL models have significant uncertainties in their SWE estimations over the low-elevation areas. Comparing the L1 norm between climatology estimations (the subtitles in Figure 11), the LSTM L1 norm is smaller than those for the reference datasets, i.e., $L1(\mathrm{UCLA}, \mathrm{UA})$ is the largest for both whole-area and high-elevation climatology estimations. This suggests the LSTM model is always in the uncertainty range of the selected process-based datasets, which provides evidence of its credibility.

In addition to differences in peak SWE, these models/datasets yield different accumulation and melt dates. Both the UA dataset and the LSTM model exhibit earlier accumulation (melt) dates when the SWE starts increasing (hits zero) in comparison to the UCLA dataset. Unlike the magnitude bias, this difference persists as elevation increases. We are not able to determine which model/dataset generates the more precise melt date. A further evaluation is needed to draw such conclusions, which is out of our scope here.

5 Discussion

One question that arises in this study relates to the importance of horizontal resolution for the accuracy of the SWE prediction. Obviously, higher resolution alone should not be conflated with higher performance – but in mountainous regions, where topography and solar insolation can vary rapidly over short distances, the resolution is important to properly capture SWE daily-to-seasonal cycles. However, significant uncertainties in snow products persist over short distances, which are exemplified by a relative performance at SNOTEL station locations. In table 1, the UCLA product was first regridded to the 4km UA grid, then interpolated to SNOTEL station locations for comparison, yielding a median NSE of 0.708. However, directly regridding the UCLA product to the SNOTEL station location, which one might expect would be far more accurate because of the finer grid spacing of the UCLA product, yields an even lower median NSE of 0.641 when assessed over the whole period. While this difference is likely to be primarily driven by observational uncertainty in SWE, we postulate that there may be another factor in play: specifically, given the significant differences in snow dynamics over relatively short spatial distances, it may be the case that accumulating SWE over a coarse grid cell may mute sharp variations in the spatial character of SWE and so could match more closely to the SNOTEL station. This is also corroborated by the spatial variability of UCLA SWE estimations within the UA grid boxes, as illustrated in Figure S8.

Because of the relatively fine scale of mountainous features, it is also the case that high-resolution static inputs do not necessarily yield better performance. The impact of static feature resolution is investigated with the PRISM 800-meter topographic data (Daly et al., 2008). This data was used to derive elevation, TRASP and DAH for the LSTM model, and compared with TRASP and DAH inputs from 30-m USGS DEM data. At 800-m spatial resolution, the derived slope and aspect are unlikely to represent the slope and aspect at the SNOTEL station, and consequently may be invalid for use in DL models of SWE. However, the performance of the LSTM model with coarse TRASP and DAH actually increases slightly: from a median NSE of 0.901 (30-m DEM) to 0.911 (800-m DEM). This change is nonetheless significant under Mood's median test (p-value equals 0.019). This increase in performance suggests that, at least for these features, DL models do not explicitly require precise topographical features for SWE prediction. This result is again likely because the significant spatial heterogeneity of mountainous regions at finer spatial scale makes it difficult to extract a clear signal from the noise.

In this study, nearest neighbor interpolation has been applied for both in-situ and gridded product evaluation. However, it could be that this interpolation method is responsible for degrading model performance. This interpolation error is now investigated when the LSTM model is applied to develop the gridded Rocky Mountain SWE product. 105 SNOTEL stations inside the Rocky Mountains are selected, and four predictions with different interpolation processes are assessed: in-situ LSTM predictions ('LSTM-in-situ'), reference datasets (either 'UA' or 'UCLA'), SWE fraction from LSTM ('LSTM-extra-FRAC') and SWE depth from LSTM and reference dataset ('LSTM-extra-SWE-REF'). All the LSTM predictions are generated from the model presented in section 4, using the SWE fraction as the target. 'LSTM-in-situ' uses the historical maximum SWE at each SNOTEL station to transform back to the SWE depth, using the nearest gridMET forcing for each SNOTEL station. Both the 'LSTM-extra-FRAC' and 'LSTM-extra-SWE-REF' are estimates where the model is run at each UA grid point using the nearest gridMET forcing and the nearest grid point to the SNOTEL station is then selected for evaluation. 'LSTM-extra-fraction' uses LSTM-generated SWE fraction, with evaluation performed on SWE fraction from each SNOTEL station, whereas 'LSTM-extra-SWE-REF' incorporates the historical maximum SWE from reference datasets. As shown in Table 4, LSTM SWE generated at UA grid points leads to a significant NSE drop of 0.05-0.06 (when predicting fraction) or 0.07-0.10 (when predicting absolute SWE) and a corresponding increase in MAE. This suggests that significant errors emerge in the gen-

Table 4. Tabulated model comparison for prediction of SWE at SNOTEL stations in the Rocky Mountains using different interpolation methods. The top table shows performance scores on dates when observed SWE is greater than zero, while the bottom table shows the whole evaluation period (from 2008 to 2018). The best scores for each metric are shown in bold font.

Nonzero SWE	Median NSE	Median MAE (mm)	Median RMSE (mm)
LSTM-in-situ	**0.802**	**41.88**	**58.28**
UA	0.775	45.09	65.44
UCLA	0.560	70.38	98.82
LSTM-extra-FRAC	0.743	-	-
LSTM-extra-SWE-UA	0.700	57.60	77.56
LSTM-extra-SWE-UCLA	0.600	63.71	84.88
Whole Period	**Median NSE**	**Median MAE (mm)**	**Median RMSE (mm)**
LSTM-in-situ	**0.891**	**25.54**	**46.32**
UA	0.861	25.66	49.50
UCLA	0.727	47.23	78.95
LSTM-extra-FRAC	0.843	-	-
LSTM-extra-SWE-UA	0.824	35.63	60.76
LSTM-extra-SWE-UCLA	0.772	39.59	68.00

Data Evaluated	
LSTM-in-situ	gridMET nearest to SNOTEL stations, max SWE from SNOTEL stations
LSTM-extra-FRAC	gridMET nearest to UA grid points
LSTM-extra-SWE-REF	gridMET nearest to REF grid points, max SWE from reference dataset

eration of gridded SWE products when interpolating quantities among grids, even those at a similar resolution, and so interpolation should be performed sparingly. Note that the UA performance in this test is unsurprising since SNOTEL data is directly assimilated into the UA product, and so the nearest UA data point is likely to be strongly coupled to the SNOTEL station.

As seen in Table 1, all the DL models obtain higher NSE values than the SNOW-17 model. In Section 3.3, it was shown that the DL models benefit from training with many SNOTEL stations; however, the SNOW-17 model, which is tuned separately for each station, relies on a set of equations to prescribe the relevant physical processes and a much smaller set of tuning parameters. The difference in performance does not appear to arise from limited inputs: although the SNOW-17 model only takes precipitation, temperature, latitude and elevation as inputs, as shown in Table 3, a DL-reduced model with only these inputs (and others permuted) still yields a whole period median NSE of 0.874 versus SNOW-17's 0.722. Of course, in the permuted model the additional variables are not removed, only permuted, and so a fair comparison would require us to train an additional model that only uses these inputs. Doing so yields a whole period median NSE of 0.846, higher than SNOW-17 but lower than the full model. This result suggests there is still substantial room for improvement in the SNOW-17 model, although we do acknowledge that more modern process-based models are likely to yield better performance. A concerted effort to replace individual processes with SNOW-17 with data-driven models could pinpoint areas of particular deficiency, though such work is beyond the scope of this study, but an important point in how ML can also inform physics-based model development.

Although the extrapolated SWE estimations generated with our fractional SWE model require the use of maximum historical SWE from a reference dataset to obtain SWE magnitude, fractional values can still provide valuable insights via some metrics, such as snow onset and melt date (see Rhoades, Jones, and Ullrich (2018)). Consequently, one could use these metrics to quantify some features of snowpack response under climate change. While some past efforts have sought to address climate change impact on snowpack with climate model simulations, the grid spacing employed in climate models is relatively generally coarse (e.g., 28 km in Rhoades, Ullrich, and Zarzycki (2018)), and so is largely unable to capture the most rugged topography and shadow casting portions of mountainous areas and their influence on the local meteorology. Given the development of downscaled climate simulations (e.g., 1/16-degree LOCA dataset by (Pierce et al., 2014) and 4-km MACA dataset by Abatzoglou and Brown (2012)), our DL models could be used for SWE ensemble projections at much higher spatial resolution. Since the frigid temperatures of high-elevation regions provide a buffer against climate change, it is essential that SWE models operate at spatial scales fine enough to resolve mountain peaks. The necessary and sufficient spatial resolution, which is likely mountain range dependent, to get convergence in mountain range or basin-average peak SWE could be investigated in future work.

6 Conclusions

Previous studies have investigated and demonstrated that DL models are useful for Earth system applications. The present work investigates three DL models for SWE prediction over the Western US, with a focus on the Rocky Mountain region. The LSTM model, which is particularly well-suited to time-series tasks, achieves the best accuracy for SWE prediction in our experiments. The TCNN, another DL model, mimics the temporal dependency with stacked 1-D CNN layers, but without inherent states like LSTM model, its performance was somewhat worse. Attention models are also promising DL methods and have become widespread in their use for time-series tasks, especially for natural language processing (NLP). Despite also demonstrating some capacity for predicting SWE, results from the Attention model were similar to the TCNN. Besides these typical DL models, there have been efforts to combine different types of sequential layers or blocks in a hybrid model, as shown in Xu et al. (2020) and Y. Chen et al. (2020). Alternate architectures of DL models, including hybrid forms of the models discussed above, hold promise for further improvements, but are left for investigation in future work.

Compared with the SNOW-17 process-based model and select reference datasets, DL models can achieve higher accuracy (in terms of NSE), when estimating in-situ SNOTEL observations. By leveraging acceleration from GPUs, the DL model training time is reasonable and the inference is fast. Given the important role that SWE has in the mountainous hydrological cycle, DL models show promise for use in operational forecasting and long-term projection. The computational speed of DL models also allows one to generate an ensemble of SWE predictions through perturbations of the initial weights of the model, enabling probabilistic SWE predictions.

A permutation-based method is used to better interpret the proposed DL model. Precipitation and elevation are shown to be the two dominant variables for SWE prediction, consistent with our physical understanding of snowpack dynamics. Although this result is intuitive, this analysis is useful for building trust in the 'black box' ML-based model before employing it for real applications. We caution that any conclusions drawn from this interpretation could be sensitive to strong correlations among input variables. In future work, we would like to examine methods that could eliminate these input correlations. For example, one could reconstruct a set of orthogonal input variables from the original inputs using principal component analysis. These orthogonal variables would contain the same information as the original inputs, which preserves the accuracy of the DL models, and the orthogonality would simplify the interpretation process. Nonetheless, the interpretation will be drawn from the reconstructed variables, which are linear combinations of original inputs and may not represent any real physical features.

Although in-situ estimates of SWE are useful for particular applications, spatiotemporally continuous SWE predictions are needed for a wider range of applications. As a proof of concept, we apply the trained DL model to generate a gridded SWE estimation across the Rocky Mountains. A major constraint for our DL model is that most in-situ estimates of SWE are provided at mid-to-high elevations at discrete points throughout the Rocky Mountains. Therefore, the extrapolation problem for our DL model is particularly pronounced when we apply our model to a wider spatial area where the statistical properties learned from the in-situ measurements might not hold (e.g., lower elevations). Without additional training data, our extrapolation results prove that we can generalize the DL models by altering the prediction from an absolute SWE depth to its seasonality. With this transformation, the target prediction becomes an elevation-invariant quantity that can be generalized to low-elevation areas, an approach also used for climate model emulation in Beucler, Pritchard, Yuval, et al. (2021). To overcome the extrapolation problem without any loss of information (or transformation), the DL models would either need more training data in low-elevation areas (e.g., satellite images) or incorporate physical constraints into their architectures (Kashinath et al., 2021).

A limitation of our study is that it mainly focuses on the use of data-driven models and does not incorporate physical constraints. One opportunity for future work would be to add mass balance into the model, as with the model described in Hoedt et al. (2021). These physical constraints could improve the physical interpretability of these models, as well. It should be noted that although we used the UA and UCLA datasets as references for extrapolation, their accuracy cannot be directly evaluated. Indeed, differences between these two datasets are observed from both grid point-wise NSE values and area-mean climatology time series over the Rocky Mountains, indicative of the uncertainties in these datasets. Additionally, the UCLA dataset provides not only the mean SWE estimations, but also other statistics (such as median and quantiles). With DL models, such distributions could be generated along with point estimations, which would allow for the quantification of uncertainty and variability, which is useful for applications such as Earth system model development. Finally, it is clear that the mean squared error-based loss function employed in DL model training often underestimates extreme values. Generative adversarial models now being explored for Earth system modeling (Manepalli et al., 2019; Pan et al., 2021) could allow for extremes to be better captured using both sequential models and adversarial loss.

Appendix A SNOW-17 Parameters

The following parameters are tuned for the SNOW-17 model. The snow-rain partition uses a linear transition scheme, which involves PXTEMP1 and PXTEMP2, while PXTEMP is only used for the rain temperature for the energy budget.

Appendix B Hyperparameter search

Hyperparameters are set based on a grid search over a range of parameter values. The search space for these values is provided here. Each candidate model is trained with the training data and evaluated with the validation period. The model results in the best NSE value is taken as the optimal hyperparameter setting. For the Attention model, it is required that the embedding size should be divisible by the number of heads. So in the following grid search, embedding size is the product of embedding size ratio and number of heads.

Table A1. SNOW-17 parameters

Parameter	Description	Unit	Range
SCF	Gage catch deficit multiplying factor		0.9-1.2
MFMAX	Maximum melt factor during non-rain periods	$\mathrm{mm} \cdot {}^{\circ}\mathrm{C} \cdot 6\mathrm{hr}^{-1}$	0.5-1.3
MFMIN	Minimum melt factor during non-rain periods	$\mathrm{mm} \cdot {}^{\circ}\mathrm{C} \cdot 6\mathrm{hr}^{-1}$	0.1-0.6
UADJ	Average wind function during rain-on-snow periods	$\mathrm{mm} \cdot \mathrm{mb}^{-1}$	0.05-0.2
PXTEMP	Temperature that separates rain and snow	°C	0.0-2.0
PXTEMP1	Lower limit temperature dividing transition from snow	°C	-2.0-0.0
PXTEMP2	Upper limit temperature dividing rain from transition	°C	0.0-4.0

Table B1. Hyperparameter search candidates for all DL models.

LSTM					
	Hidden states	64	128	256	512
TCNN					
	Blocks	4	5	6	
	Kernel size	7	9		
	Number of kernels	16	32	64	
Attention					
	Heads	8	16		
	Embedding size ratio	1	2		
	Attention layers	2	3	4	
	Forward dimension	16	32	64	

We further tested several Attention models with 32 heads. The models generally have similar performance to the 16-head models, but take a much longer time for training. With such a small increment in performance, we decided to stop searching at 16 heads.

Acknowledgments

This study was funded by the Director, Office of Science, Office of Biological and Environmental Research of the U.S. Department of Energy Regional and Global Climate Modeling Program (RGCM) "the Calibrated and Systematic Characterization, Attribution and Detection of Extremes (CASCADE)" Science Focus Area (award no. DE-AC02-05CH11231) and the "An Integrated Evaluation of the Simulated Hydroclimate System of the Continental US" project (award no. DE-SC0016605). We would like to thank MetroIT at UC Davis for the Tempest GPU cluster. We also thank the Computational and Information System Lab for access to the Casper cluster through the Advanced Study Program at the National Center for Atmospheric Research (NCAR). We acknowledge the helpful discussion with Chaopeng Shen, Wen-Ping Tsai from Pennsylvania State University, David John Gagne from NCAR, and Chris Paciorek from Lawrence Berkeley National Laboratory.

The deep learning model code can be found at `https://github.com/ShihengDuan/code-SWE`. Our predictions for SNOTEL stations, extrapolation over the Rocky Mountains along with the necessary code can be accessed at `DOI10.5281/zenodo.6419931` (S. Duan et al., 2022). SNOTEL SWE observations can be accessed at `https://www.nrcs.usda.gov/wps/portal/wcc/home/`. GridMET atmospheric data is available at `https://`

www.climatologylab.org/gridmet.html. PRISM dataset is provided by PRISM Climate Group at: https://prism.oregonstate.edu. DEM data is provided by USGS can available through Microsoft Planetary Computer at https://planetarycomputer.microsoft.com.

References

Abatzoglou, J. T. (2013). Development of gridded surface meteorological data for ecological applications and modelling. *International Journal of Climatology*, *33*(1), 121–131.

Abatzoglou, J. T., & Brown, T. J. (2012). A comparison of statistical downscaling methods suited for wildfire applications. *International journal of climatology*, *32*(5), 772–780.

Addor, N., Newman, A. J., Mizukami, N., & Clark, M. P. (2017). The camels data set: catchment attributes and meteorology for large-sample studies. *Hydrology and Earth System Sciences*, *21*(10), 5293–5313.

Anderson, E. (2006). *Snow accumulation and ablation model – snow-17.* (https://www.weather.gov/media/owp/oh/hrl/docs/22snow17.pdf)

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 ....

Anderson, E. A. (1976). *A point energy and mass balance model of a snow cover.* Stanford University.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*.

Balestriero, R., Pesenti, J., & LeCun, Y. (2021). Learning in high dimension always amounts to extrapolation. *arXiv preprint arXiv:2110.09485*.

Barnes, E. A., Toms, B., Hurrell, J. W., Ebert-Uphoff, I., Anderson, C., & Anderson, D. (2020). Indicator patterns of forced change learned by an artificial neural network. *Journal of Advances in Modeling Earth Systems*, *12*(9), e2020MS002195.

Berghuijs, W. R., Harrigan, S., Molnar, P., Slater, L. J., & Kirchner, J. W. (2019). The relative importance of different flood-generating mechanisms across europe. *Water Resources Research*, *55*(6), 4582–4593.

Beucler, T., Pritchard, M., Rasp, S., Ott, J., Baldi, P., & Gentine, P. (2021). Enforcing analytic constraints in neural networks emulating physical systems. *Physical Review Letters*, *126*(9), 098302.

Beucler, T., Pritchard, M., Yuval, J., Gupta, A., Peng, L., Rasp, S., ... others

(2021). Climate-invariant machine learning. *arXiv preprint arXiv:2112.08440*.

Blöschl, G. (1999). Scaling issues in snow hydrology. *Hydrological processes*, *13*(14-15), 2149–2175.

Böhner, J., & Antonić, O. (2009). Land-surface parameters specific to topo-climatology. *Developments in soil science*, *33*, 195–226.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123–140.

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.

Chen, X., Leung, L. R., Gao, Y., & Liu, Y. (2021). Response of us west coast mountain snowpack to local sea surface temperature perturbations: insights from numerical modeling and machine learning. *Journal of Hydrometeorology*, *22*(4), 1045–1062.

Chen, Y., Dai, X., Liu, M., Chen, D., Yuan, L., & Liu, Z. (2020). Dynamic convolution: Attention over convolution kernels. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 11030–11039).

Civitarese, D. S., Szwarcman, D., Zadrozny, B., & Watson, C. (2021). Extreme precipitation seasonal forecast using a transformer neural network. *arXiv preprint arXiv:2107.06846*.

Cristea, N. C., Breckheimer, I., Raleigh, M. S., HilleRisLambers, J., & Lundquist, J. D. (2017). An evaluation of terrain-based downscaling of fractional snow covered area data sets based on lidar-derived snow data and orthoimagery. *Water Resources Research*, *53*(8), 6802–6820.

Daly, C., Halbleib, M., Smith, J. I., Gibson, W. P., Doggett, M. K., Taylor, G. H., ... Pasteris, P. P. (2008). Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous united states. *International Journal of Climatology: a Journal of the Royal Meteorological Society*, *28*(15), 2031–2064.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Duan, Q., Gupta, V. K., & Sorooshian, S. (1993). Shuffled complex evolution approach for effective and efficient global minimization. *Journal of optimization theory and applications*, *76*(3), 501–521.

Duan, Q., Sorooshian, S., & Gupta, V. K. (1994). Optimal use of the sce-ua global optimization method for calibrating watershed models. *Journal of hydrology*, *158*(3-4), 265–284.

Duan, S., Ullrich, P., Risser, M., & Rhoades, A. (2022). *Historical daily snow water equivalent (swe) estimations over the western us and the rocky mountains.* Zenodo. Retrieved from `https://zenodo.org/record/6419930` doi: 10.5281/ZENODO.6419930

Duan, S., Ullrich, P., & Shu, L. (2020). Using convolutional neural networks for streamflow projection in california. *Frontiers in Water*, *2*, 28.

Fang, K., Kifer, D., Lawson, K., Feng, D., & Shen, C. (2022). The data synergy effects of time-series deep learning models in hydrology. *Water Resources Research*, *58*(4), e2021WR029583.

Fang, Y., Liu, Y., & Margulis, S. A. (2022). A western united states snow reanalysis dataset over the landsat era from water years 1985 to 2021. *Scientific Data*, *9*(1), 677.

Feng, D., Beck, H., Lawson, K., & Shen, C. (2022). The suitability of differentiable, learnable hydrologic models for ungauged regions and climate change impact assessment. *Hydrology and Earth System Sciences Discussions*, 1–28.

Feng, D., Fang, K., & Shen, C. (2020). Enhancing streamflow forecast and extracting insights using long-short term memory networks with data integration at continental scales. *Water Resources Research*, *56*(9), e2019WR026793.

Gagne II, D. J., Haupt, S. E., Nychka, D. W., & Thompson, G. (2019). Interpretable deep learning for spatial analysis of severe hailstorms. *Monthly Weather Review*, *147*(8), 2827–2845.
Hatchett, B. J., Rhoades, A. M., & McEvoy, D. J. (2022). Monitoring the daily evolution and extent of snow drought. *Natural Hazards and Earth System Sciences*, *22*(3), 869–890.
He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).
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.
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).
Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.
Hoedt, P.-J., Kratzert, F., Klotz, D., Halmich, C., Holzleitner, M., Nearing, G., … Klambauer, G. (2021). Mc-lstm: Mass-conserving lstm. *arXiv preprint arXiv:2101.05186*.
Jennings, K., Winchell, T. S., Livneh, B., & Molotch, N. P. (2018). Spatial variation of the rain–snow temperature threshold across the Northern Hemisphere. *Nature Communications*, *9*(1). doi: 10.1038/s41467-018-03629-7
Kashinath, K., Mustafa, M., Albert, A., Wu, J., Jiang, C., Esmaeilzadeh, S., … others (2021). Physics-informed machine learning: case studies for weather and climate modelling. *Philosophical Transactions of the Royal Society A*, *379*(2194), 20200093.
King, F., Erler, A. R., Frey, S. K., & Fletcher, C. G. (2020). Application of machine learning techniques for regional bias correction of snow water equivalent estimates in ontario, canada. *Hydrology and Earth System Sciences*, *24*(10), 4887–4902.
Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019a). Benchmarking a catchment-aware long short-term memory network (lstm) for large-scale hydrological modeling. *arXiv preprint arXiv:1907.08456*.
Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019b). 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.
Lea, C., Flynn, M. D., Vidal, R., Reiter, A., & Hager, G. D. (2017). Temporal convolutional networks for action segmentation and detection. In *proceedings of the ieee conference on computer vision and pattern recognition* (pp. 156–165).
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 lstm-based models with four lumped conceptual models. *Hydrology and Earth System Sciences Discussions*, 1–41.
Lin, T., Wang, Y., Liu, X., & Qiu, X. (2021). A survey of transformers. *arXiv preprint arXiv:2106.04554*.
Livneh, B., & Badger, A. M. (2020). Drought less predictable under declining future snowpack. *Nature Climate Change*, *10*(5), 452–458.
Livneh, B., Bohn, T. J., Pierce, D. W., Munoz-Arriola, F., Nijssen, B., Vose, R., … Brekke, L. (2015). A spatially comprehensive, hydrometeorological data set for mexico, the us, and southern canada 1950–2013. *Scientific data*, *2*(1), 1–12.

Manepalli, A., Albert, A., Rhoades, A., Feldman, D., & Jones, A. D. (2019). Emulating numeric hydroclimate models with physics-informed cgans. In *Agu fall meeting 2019.*

Margulis, S. A., Liu, Y., & Baldo, E. (2019). A joint landsat-and modis-based reanalysis approach for midlatitude montane seasonal snow characterization. *Frontiers in Earth Science*, *7*, 272.

McCrary, R. R., McGinnis, S., & Mearns, L. O. (2017). Evaluation of snow water equivalent in narccap simulations, including measures of observational uncertainty. *Journal of Hydrometeorology*, *18*(9), 2425 - 2452. Retrieved from `https://journals.ametsoc.org/view/journals/hydr/18/9/jhm-d-16-0264_1.xml` doi: https://doi.org/10.1175/JHM-D-16-0264.1

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.

Meyal, A. Y., Versteeg, R., Alper, E., Johnson, D., Rodzianko, A., Franklin, M., & Wainwright, H. (2020). Automated cloud based long short-term memory neural network based swe prediction. *Frontiers in Water*, *2*.

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.

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.

Odry, J., Boucher, M., Cantet, P., Lachance-Cloutier, S., Turcotte, R., & St-Louis, P. (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.

Pan, B., Anderson, G. J., Goncalves, A., Lucas, D. D., Bonfils, C. J., Lee, J., ... Ma, H.-Y. (2021). Learning to correct climate projection biases. *Journal of Advances in Modeling Earth Systems*, *13*(10), e2021MS002509.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32* (pp. 8024–8035). Curran Associates, Inc. Retrieved from `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`

Pierce, D. W., Cayan, D. R., & Thrasher, B. L. (2014). Statistical downscaling using localized constructed analogs (loca). *Journal of Hydrometeorology*, *15*(6), 2558–2585.

Raleigh, M. S., & Lundquist, J. D. (2012). Comparing and combining swe estimates from the snow-17 model using prism and swe reconstruction. *Water Resources Research*, *48*(1).

Rhoades, A. M., Hatchett, B. J., Risser, M. D., Collins, W. D., Bambach, N. E., Huning, L. S., ... others (2022). Asymmetric emergence of low-to-no snow in the midlatitudes of the american cordillera. *Nature Climate Change*, 1–9.

Rhoades, A. M., Jones, A. D., & Ullrich, P. A. (2018). Assessing mountains as natural reservoirs with a multimetric framework. *Earth's Future*, *6*(9), 1221–1241.

Rhoades, A. M., Ullrich, P. A., & Zarzycki, C. M. (2018). Projecting 21st century snowpack trends in western usa mountains using variable-resolution cesm. *Climate Dynamics*, *50*(1), 261–288.

Roberts, D. W., & Cooper, S. V. (1989). Concepts and techniques of vegetation mapping. *Land classifications based on vegetation: applications for resource management*, 90–96.

Serreze, M. C., Clark, M. P., Armstrong, R. L., McGinnis, D. A., & Pulwarty, R. S. (1999). Characteristics of the western united states snowpack from snowpack telemetry (snotel) data. *Water Resources Research*, *35*(7), 2145–2160.

Siirila-Woodburn, E. R., Rhoades, A. M., Hatchett, B. J., Huning, L. S., Szinai, J., Tague, C., ... others (2021). A low-to-no snow future and its impacts on water resources in the western united states. *Nature Reviews Earth & Environment*, *2*(11), 800–819.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

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.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1–9).

Tay, Y., Dehghani, M., Bahri, D., & Metzler, D. (2020). Efficient transformers: A survey. *arXiv preprint arXiv:2009.06732*.

Toms, B. A., Barnes, E. A., & Ebert-Uphoff, I. (2020). Physically interpretable neural networks for the geosciences: Applications to earth system variability. *Journal of Advances in Modeling Earth Systems*, *12*(9), e2019MS002002.

Tsymbal, A. (2004). The problem of concept drift: definitions and related work. *Computer Science Department, Trinity College Dublin*, *106*(2), 58.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Wang, X., Kondratyuk, D., Christiansen, E., Kitani, K. M., Alon, Y., & Eban, E. (2021). *Wisdom of committees: An overlooked approach to faster and more accurate models.*

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*(3), e2021WR031033.

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.

Xiang, Z., Yan, J., & Demir, I. (2020). A rainfall-runoff model with lstm-based sequence-to-sequence learning. *Water resources research*, *56*(1), e2019WR025326.

Xu, G., Ren, T., Chen, Y., & Che, W. (2020). A one-dimensional cnn-lstm model for epileptic seizure recognition using eeg signal analysis. *Frontiers in Neuroscience*, *14*, 1253.

Yan, J., Mu, L., Wang, L., Ranjan, R., & Zomaya, A. Y. (2020). Temporal convolutional networks for the advance prediction of enso. *Scientific reports*, *10*(1), 1–15.

Ye, F., Hu, J., Huang, T.-Q., You, L.-J., Weng, B., & Gao, J.-Y. (2021). Transformer for ei niño-southern oscillation prediction. *IEEE Geoscience and Remote Sensing Letters*.

Zeng, X., Broxton, P., & Dawson, N. (2018). Snowpack change from 1982 to 2016 over conterminous united states. *Geophysical Research Letters*, *45*(23), 12–940.