LSTM-Based Data Integration to Improve Snow Water Equivalent Prediction and Diagnose Error Sources

YALAN SONG,[a] WEN-PING TSAI,[a,b] JONAH GLUCK,[c] ALAN RHOADES,[d] COLIN ZARZYCKI,[e] RACHEL MCCRARY,[f] KATHRYN LAWSON,[a] AND CHAOPENG SHEN[a]

[a] *Civil and Environmental Engineering, The Pennsylvania State University, University Park, Pennsylvania*
[b] *Hydraulic and Ocean Engineering, National Cheng Kung University, Tainan, Taiwan*
[c] *Computer Science, Boston University, Boston, Massachusetts*
[d] *Earth and Environmental Sciences Area, Lawrence Berkeley National Laboratory, Berkeley, California*
[e] *Meteorology and Atmospheric Science, The Pennsylvania State University, University Park, Pennsylvania*
[f] *National Center for Atmospheric Research, Boulder, Colorado*

(Manuscript received 1 December 2022, in final form 10 October 2023, accepted 11 October 2023)

ABSTRACT: Accurate prediction of snow water equivalent (SWE) can be valuable for water resource managers. Recently, deep learning methods such as long short-term memory (LSTM) have exhibited high accuracy in simulating hydrologic variables and can integrate lagged observations to improve prediction, but their benefits were not clear for SWE simulations. Here we tested an LSTM network with data integration (DI) for SWE in the western United States to integrate 30-day-lagged or 7-day-lagged observations of either SWE or satellite-observed snow cover fraction (SCF) to improve future predictions. SCF proved beneficial only for shallow-snow sites during snowmelt, while lagged SWE integration significantly improved prediction accuracy for both shallow- and deep-snow sites. The median Nash–Sutcliffe model efficiency coefficient (NSE) in temporal testing improved from 0.92 to 0.97 with 30-day-lagged SWE integration, and root-mean-square error (RMSE) and the difference between estimated and observed peak SWE values $d_{\max}$ were reduced by 41% and 57%, respectively. DI effectively mitigated accumulated model and forcing errors that would otherwise be persistent. Moreover, by applying DI to different observations (30-day-lagged, 7-day-lagged), we revealed the spatial distribution of errors with different persistent lengths. For example, integrating 30-day-lagged SWE was ineffective for ephemeral snow sites in the southwestern United States, but significantly reduced monthly-scale biases for regions with stable seasonal snowpack such as high-elevation sites in California. These biases are likely attributable to large interannual variability in snowfall or site-specific snow redistribution patterns that can accumulate to impactful levels over time for nonephemeral sites. These results set up benchmark levels and provide guidance for future model improvement strategies.

KEYWORDS: Snowpack; Hydrology; Machine learning; Deep learning; Error analysis; Snow

1. Introduction

Snowpack is a critical source of water supply in many parts of the world. In the western United States, snowmelt accounts for more than half of the total runoff (Li et al. 2017) and is a major water source for agriculture, human consumption, and hydroelectric power generation (Kopytkovskiy et al. 2015; Magnusson et al. 2020; Qin et al. 2020; Siirila-Woodburn et al. 2021; Ullrich et al. 2018; Vano 2020). Snowmelt timing is also important for estimating peak flows and nutrient exports (Corriveau et al. 2011, 2013). Water managers in the snow-dominated western United States rely on the snow water equivalent (SWE) measurements at ~800 Snowpack Telemetry (SNOTEL) sites to monitor snow drought and make water management decisions (Hatchett et al. 2022; Nowak et al. 2021; USDA 2022). In situ SWE data, along with snow depth data, have also been utilized to produce interpolated SWE maps (Broxton et al. 2016; Dawson et al. 2018; NOHRSC 2004). If SWE values at in situ locations can be accurately predicted, they could be of significant value to data users and stakeholders. A continuous, high-quality dataset without gaps could be particularly useful for studying the snowpack variability and trends over long periods of time at sites that do not have seamless daily data records for analysis.

SWE prediction is often achieved by combining observational data with a model using a data assimilation (DA) approach like ensemble Kalman filtering (EnKF) (Diro and Lin 2020; King et al. 2020; Leisenring and Moradkhani 2011; Slater and Clark 2006). Slater and Clark (2006) improved SWE estimations by assimilating ensemble forcing data and observations of SWE into a conceptual snow model (SNOW-17). DA updates the internal states of a process-based model using covariance-based, variational, particle swarm, or other methods, and the updated model can provide better predictions of model outputs. The performance of DA depends on the validity of assumptions such as linearity for covariance-based methods and the probability distributions for variational methods. It can also be impacted by the realism of the underlying physical model and choices such as composition of the covariance matrix and bias correction strategies (Evensen 2009).

Recently, machine learning approaches such as long short-term memory (LSTM) deep learning (DL) networks have shown great promise across many applications. LSTM models have demonstrated high performance for modeling soil moisture (Fang et al. 2017; Fang and Shen 2020; Li et al. 2021; O and Orth 2021), streamflow (Feng et al. 2020, 2021; Kratzert et al. 2019; Ma et al.

Corresponding author: Chaopeng Shen, cshen@engr.psu.edu

DOI: 10.1175/JHM-D-22-0220.1

2021; Ouyang et al. 2021; Xiang et al. 2020), stream temperature (Rahmani et al. 2021a,b), dissolved oxygen (Zhi et al. 2021, 2023), and groundwater (Afzaal et al. 2020), among others (Shen and Lawson 2021). Furthermore, these models offer uncertainty (Fang et al. 2020; Klotz et al. 2022) and physical parameter (Tsai et al. 2021; Feng et al. 2022; Shen et al. 2023; Aboelyazeed et al. 2023) estimation functionalities.

In the context of making forecasts with LSTM-based models, one can either use data integration (DI) (Feng et al. 2020) (or "autoregression"), where recent observations are included as inputs so that LSTM learns how to best make use of such information, or variational data assimilation (Nearing et al. 2022), where an update to the LSTM internal states is applied to minimize the difference between simulation and observation. Neither of the two options directly updates the observed variable: if we have an LSTM model that simulates SWE, we cannot simply update SWE because in LSTM models, SWE is merely an output (or "display variable") of the network and not the state variable involved in calculating such an output; thus updating SWE would have no impact on subsequent predictions. One could instead use the second option where LSTM's internal states are updated using variational data assimilation. However, this approach can be expensive and, based on previous evaluation, may not be as optimal as DI (Nearing et al. 2022). LSTM deep networks showed high flexibility and performance in leveraging recent observations to improve prediction with DI. Fang and Shen (2020) were able to substantially improve soil moisture forecasts over the conterminous United States (CONUS). They reduced the 3-day-ahead prediction root-mean-square error (RMSE) to 0.022, which was far lower than previously published estimates (Koster et al. 2017). Feng et al. (2020) showed DI improved 1-day-ahead median Nash–Sutcliffe model efficiency coefficient (NSE) values for streamflow from 0.74 to 0.86 for a benchmark dataset over the CONUS. DI lets the neural network decide how to best fuse information from the lagged observations with other inputs to influence the output. DI mostly corrects the errors accumulated in the forcing data or in the model dynamics, and prevents such errors from influencing the ensuing simulations (Fang and Shen 2020). Due to the high autocorrelation of SWE time series, it is expected that SWE prediction should benefit significantly from DI (Fang and Shen 2020). However, as SWE is determined by a combination of antecedent conditions and can be impacted by complex processes (e.g., rain-on-snow events), it was uncertain if DI's benefit was strong at the weekly or monthly scales, and, if so, in which locations it would be most beneficial. For SWE modeling, Meyal et al. (2020) trained LSTM models on five SNOTEL sites and reported high forecast performance using the most recent SWE observations. However, they used only a small fraction of available SNOTEL stations, which likely limits the usefulness of interpolation or extrapolation of these predictions for other sites (Fang et al. 2022).

While the in situ SWE network is valuable, the sites are unevenly distributed in space. Satellite data can observe snow cover fraction (SCF) over the entire surface of the globe at high spatiotemporal resolution, but SCF does not readily translate into SWE as the relationship is complex, hysteretic, and nonlinear (Egli and Jonas 2009; Luce and Tarboton 2004; Magand et al. 2014). There are also airborne sensing data (Painter et al. 2016) but they have not been systematically collected for a long period of time. Previous efforts have attempted to utilize SCF in multisensor data assimilation (Girotto et al. 2020) but have had limited success. Given the demonstrated success of deep learning, it is worthwhile to see if DL models can make better use of this data.

Here we applied a DI scheme for SWE simulations at monitored sites over the western United States and compared different model formulations and lag times. We had three research questions:

1) How much improvement in SWE prediction can be obtained by integrating SWE observations with 30-day or 7-day lag times?
2) How useful are recent observations of SWE and SCF data, respectively, for sites without in situ data (unmonitored sites)?
3) What are the orders of magnitude of predictive errors for SWE, what are their characteristics (e.g., spatiotemporal error, long-term cumulative error, or short-term flash error), and which locations are most susceptible to these errors?

Our research primarily focused on the SWE prediction at monitored sites, as these sites are highly valuable for water resource managers and have ground-truth measurements for verification. We also provided high-quality predictions at unmonitored sites and worked to separate out different error sources.

2. Methods

a. Datasets

SNOTEL is a network of automated stations that monitor SWE and meteorological parameters across mountainous regions in the western United States (Serreze et al. 1999). At each SNOTEL site, a "snow pillow" uses a pressure transducer to measure the weight of the snowpack, which is then used to calculate the SWE (Gan et al. 2021). Daily SWE from 525 stations (those having data available from 2001 to 2019) was used as training and test data in this work.

We used daily snow cover fraction (SCF) with 500-m spatial resolution from the Moderate Resolution Imaging Spectroradiometer (MODIS) on board the *Terra* satellite (Hall and Riggs 2021) as potential observations to be assimilated. The specific product used was MOD10A1F, derived from the MODIS Snow Cover Daily L3 Global 500-m Grid (MOD10A1). Gap-filling techniques were employed to compensate for cloud cover or poor data quality by incorporating observations from the most recent clear-sky day. However, some gaps still persisted in the data due to continuous cloud cover, and to address this issue, we substituted the remaining no-data periods with a value of zero. MOD10A1F has been reported to have a high accuracy of 96.2% when compared with the snow cover map converted from a 3-m-resolution snow-depth product by an aerial laser scanner (Stillinger et al. 2023). For each SNOTEL site, the SCF value was calculated as the mean of the MODIS SCF data across all pixels inside the 4-km grid box around the site. The 4-km resolution was chosen because we wanted to use the same resolution as the parameters from gridMET (we did test both 500-m and 4-km

resolution and achieved similar performance but the 4-km resolution had less noise and fewer gaps due to continuous cloud cover; data not shown).

This work focused on the hydrologic component of the prediction—that is, rather than using a meteorological forecast, we used reanalysis forcings to drive the model. This allows analysis of errors in different parts of the model. Moreover, as we show, even reanalysis products (often seen as more accurate than forecasts) can introduce large errors. We used gridded surface meteorological data (gridMET) as our forcing data, a dataset of daily high spatial resolution (~4 km; 1/24°) surface reanalysis meteorological data covering the contiguous United States from 1979 through the present with 1-day resolution (Abatzoglou 2013). It blends climate data from PRISM (800-m resolution) and NLDAS-2 (12-km resolution) using climatically aided interpolation. PRISM is upscaled and NLDAS-2 is downscaled to a 4-km resolution. The gridMET variables used in the LSTM model were precipitation, 2-m surface air temperature, downward surface shortwave radiation, wind velocity, and humidity. The values of variables on the closest pixel to each SNOTEL site were employed as the forcing data. We conducted tests utilizing both the meteorological data from gridMET and SNOTEL to separate out potential meteorological forcing errors. SNOTEL provides potentially more accurate forcing data at the SNOTEL sites, but this information is not available elsewhere and thus would not work for predictions outside SNOTEL sites. Other variables such as measured or simulated soil and snow temperatures may be added, too, and could potentially have minor benefits, but we did not include them in the present study as these measurements were not available for most SNOTEL sites, and we wanted to focus on widely available observational data.

With respect to the static physiographic attributes (A), we employed a range of predictors for the main LSTM unit: (i) latitude; (ii) topographic characteristics including elevation, slope, and aspect; and (iii) land cover characteristics (dominant land cover, dominant land cover fraction, and forest fraction) from MODIS Land Cover Type Yearly L3 Global 500 m (MCD12Q1). The topographic characteristics, e.g., slope, elevation, and aspect, can affect the solar insolation on the snowpack, while the vegetation cover and forest gaps also impact snow accumulation and melt by canopy interception, wind dampening, local longwave radiation emittance, and shading of shortwave radiation (Smyth et al. 2022). Although we simply included all logically relevant and widely available static attributes as inputs, we ran experiments with different setups to show the impacts of adding or removing some attributes (Table A2 in the appendix). Additional information such as tree morphometry—for example, shape, height, crown size, and density—could have impacts on snow (Seyednasrollah and Kumar 2013). We did not include them because of large-scale availability issues, but they could be tested in the future.

b. Modeling

As a quick summary, we trained two types of models. The first type is a forward model that does not perform DI. This model can be written concisely as

$$\mathrm{SWE}^{1:t} = \mathrm{LSTM}(f^{1:t}, A), \tag{1}$$

where t is the current time step, f represents the atmospheric forcings, and A represents the static attributes that characterize a site. This formulation is a sequence-to-sequence model and information from the future is not used during prediction (Hochreiter and Schmidhuber 1997; Shen 2018; Fang and Shen 2020). The second type of model uses DI, which refers to the incorporation of recent observations into the model. It can be concisely written as

$$\mathrm{SWE}^{T+1:t} = \mathrm{LSTM}(f^{T+1:t}, A, y^{1:t-T}). \tag{2}$$

In other words, we feed a T-day lagged variable y, which could be either SWE or SCF, and let LSTM decide how to best use it to update the internal states so it could make a better prediction (Fig. 1a). There are no special steps involved to achieve the fusion of data, and all information is considered. The only difference between the DI-LSTM model and the forward model without DI is whether the lagged observations are integrated (concatenated) in the inputs.

The LSTM models were developed based on previous soil moisture and streamflow prediction work that had data integration components (Fang and Shen 2020; Feng et al. 2020). LSTM is a special recurrent neural network that not only has recurrent connections but also contains input, output, and forget gates to add or remove information from the cell state of the LSTM (Fig. 1b) (Hochreiter and Schmidhuber 1997; Graves 2012). The gates in a cell state act like filters that determine which information to remember and which to discard over long periods of time. These components were collectively designed to help address the challenge of vanishing gradients in deep learning, a problem that arises when the gradient decreases exponentially across time steps (Hochreiter et al. 2001). The LSTM network and our whole workflow were implemented in PyTorch (Paszke et al. 2019), an open source, Python-based, machine learning framework. More details about the LSTM implementation here can be found in our previous papers, Fang et al. (2017) and Fang and Shen (2020). Given inputs I (a concatenation of meteorological forcings, physiographic attributes, and lagged observations), our LSTM algorithm can be written as the following:

$$\begin{aligned}
&\text{Input transformation}: x^t = \mathrm{ReLU}(\mathbf{W}_I I^t + \mathbf{b}_I),\\
&\text{Input node}: g^t = \tanh[\mathcal{D}(\mathbf{W}_{gx} x^t) + \mathcal{D}(\mathbf{W}_{gh} h^{t-1}) + \mathbf{b}_g],\\
&\text{Input gate}: i^t = \sigma[\mathcal{D}(\mathbf{W}_{ix} x^t) + \mathcal{D}(\mathbf{W}_{ih} h^{t-1}) + \mathbf{b}_i],\\
&\text{Forget gate}: f^t = \sigma[\mathcal{D}(\mathbf{W}_{fx} x^t) + \mathcal{D}(\mathbf{W}_{fh} h^{t-1}) + \mathbf{b}_f],\\
&\text{Output gate}: o^t = \sigma[\mathcal{D}(\mathbf{W}_{ox} x^t) + \mathcal{D}(\mathbf{W}_{oh} h^{t-1}) + \mathbf{b}_o],\\
&\text{Cell state}: s^t = g^t \odot i^t + s^{t-1} \odot f^t,\\
&\text{Hidden state}: h^t = \tanh(s^t) \odot o^t, \quad \text{and}\\
&\text{Output}: y^t = \mathbf{W}_{hy} h^t + \mathbf{b}_y,
\end{aligned} \tag{3}$$

where $\mathbf{W}$ and $\mathbf{b}$ are the network weights and bias parameters, respectively; h are the hidden states; tanh is the tangent hyperbolic function used as an activation function; and $\mathcal{D}$ is the dropout operator, which is a regularization technique to prevent overfitting by randomly dropping out a proportion of

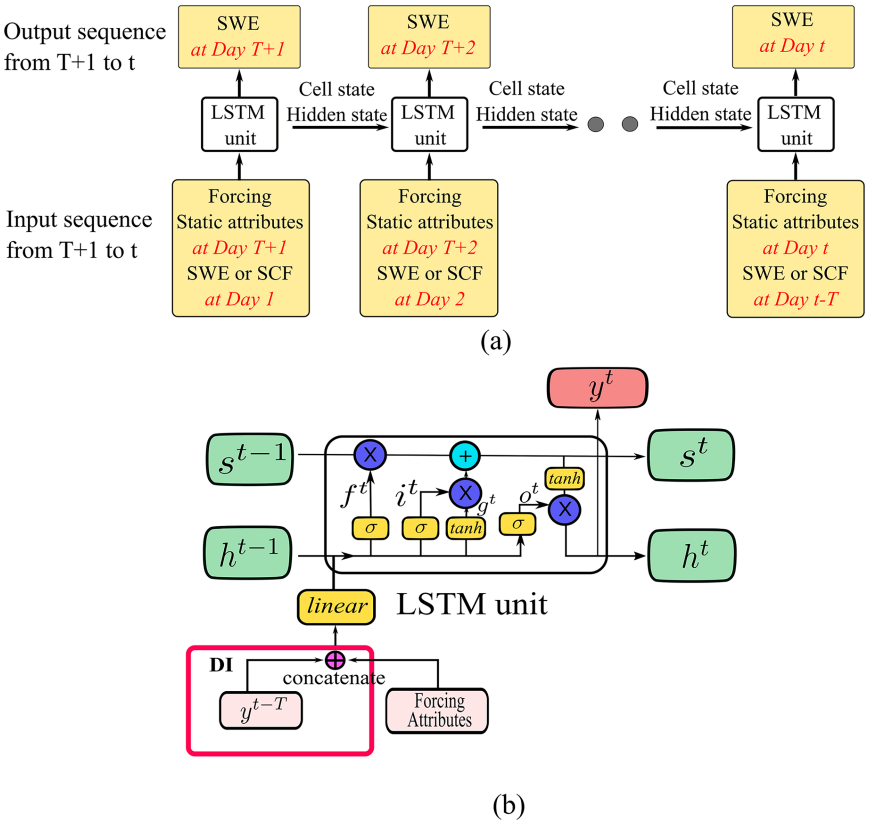


FIG. 1. LSTM model with DI: (a) the sequence-to-sequence LSTM model with data integration of T-day-lagged SWE or SCF, and (b) a schematic view of the workflow in an LSTM unit.

neurons in a neural network during training. All the variables are matrices.

DI in the framework of LSTM is straightforward—it either simply concatenates the lagged observations with other inputs to be supplied to LSTM as illustrated in Fig. 1, or sends in lagged observations via additional neuron units, which are unnecessary for this work. It is conceptually similar to autoregressive models that use past values of a variable as inputs to predict its future values, but the past variables are not limited to the variable to be predicted. In contrast with this simplistic method, conceptual or process-based models often employ data assimilation techniques to incorporate recent measurements (Houser et al. 1998; Vrugt et al. 2006; Clark et al. 2008; Nearing et al. 2022). As explored in other studies (Nearing et al. 2022), an autoregressive model outperformed a variational data assimilation approach that used a separate backward state update step. In this work, we tested our DI algorithm with two variables: lagged SWE and lagged SCF. Even though it was expected that SCF would not provide too much information about SWE when snow is deep, SCF is more widely available than SWE (as it can be observed by satellites), and we wondered if it could help improve predictions at shallow-snow sites.

To avoid overtuning hyperparameters, the hyperparameter combinations were inherited from our previous streamflow model (Feng et al. 2020): a batch size of 100, a hidden-state size of 256, and a dropout rate of 0.5. A 365-day training sequence length was needed for the forward model but a shorter length was sufficient for the DI models (Table A1). We employed 365 days for all models for simplicity, considering the period of snow dynamics is one year for ephemeral and seasonal snowpacks characteristic of the western United States. All the models were trained until reaching convergence. Each training job only needed 2.5 h on an NVIDIA TITAN Xp. The stochastic gradient descent method AdaDelta (Zeiler 2012) was used to automatically adapt the learning rate.

c. *Temporal and spatial cross-validation testing*

In the forward model, recent observations are not used; that is, there is no $y^{1:t-T}$ term in the input in Eq. (1). In the DI model, the $y^{1:t-T}$ term can be either SCF or SWE data at the site, and the lag T can be either 7 [denoted as DI(SCF, 7) and DI(SWE, 7)] or 30 [DI(SCF, 30) and DI(SWE, 30)] days. A shorter lag results in the integration of a more recent observation and, most of the time, better predictive metrics. For each model, the same integrated term and lag were employed in both training and testing.

All the models were trained from 2001 to 2015 and evaluated for temporal, spatial, and spatiotemporal performance. The temporal test was conducted using temporal cross validation, where the models were trained on all 525 SNOTEL sites from 2001 to 2015 and then tested on all sites from 2016 to

2019. The spatial test was conducted using a 10-fold spatial cross-validation approach where the 525 SNOTEL sites were randomly divided into 10 folds. The model was trained from 2001 to 2015 on 9 of the folds and tested from 2001 to 2015 on the remaining (untrained) sites. The test was run a total of 10 times while rotating the held-out fold, and thus every site became a test site once. In the spatiotemporal test, the models were trained in the same way as in the spatial test, but the evaluation was conducted in the 4 years after the training period, from 2016 to 2019.

d. Evaluation metrics

LSTM performance was evaluated over the testing years using five statistical metrics: bias, RMSE, Pearson's correlation R, NSE, and the absolute difference $d_{\max}$ between the maximum model simulation and the maximum observation over the water year. The NSE (McCuen et al. 2006) considers bias, with a perfect model yielding a value of 1, while poor performing models can have infinitely negative values. Additionally, the absolute difference between the maximum model estimate and the maximum observation over the water year, referred to as $d_{\max}$, was calculated using

$$d_{\max} = \frac{\sum_{i=1}^{n} |(\mathrm{SWE}_{i,\max} - \mathrm{SWE}^{*}_{i,\max})|}{n}, \tag{4}$$

where $\mathrm{SWE}_{i,\max}$ and $\mathrm{SWE}^{*}_{i,\max}$ are respectively the maximum model simulation and maximum observations at pixel i evaluated in the water year, and n is the number of evaluated sites.

We used the daily 4-km gridded SWE data from the University of Arizona (UA) dataset (Broxton et al. 2016) as a benchmark to evaluate our LSTM models. The UA dataset was obtained by interpolating SWE from SNOTEL stations using ordinary kriging. The SNOTEL SWE was normalized by the net accumulated snowfall that considered both accumulated snowfall and cumulative ablation. The interpolated SWE values at those unmonitored stations were then denormalized by the net accumulated snowfall, as described in previous studies (Broxton et al. 2016; Zeng et al. 2018). This dataset also incorporated snow-depth measurements from thousands of National Weather Service (NWS) Cooperative Observer Program (COOP) stations, which was achieved by the new snow density model described in Dawson et al. (2017). We compared the SWE data from the UA dataset on the nearest grid points to the SNOTEL sites with our LSTM models' predictions for the period of 2001 to 2019. The UA dataset can be accessed through the National Snow and Ice Data Center (NSIDC).

e. Model error analysis

As discussed earlier, LSTM has been shown to be skillful at capturing temporal dynamics especially for monitored sites, which means it can be used as an approximate measure of the information content of the input. To assist in interpreting the results, we classified model error into three types: 1) temporal error, produced when a model was trained on some sites in a certain period and tested on the same sites in a different period—it is introduced due to the model not being trained to respond correctly in time; 2) spatial error, produced when a model was trained on some sites in a certain period and tested on different, untrained sites in the same period—it is due either to the model not responding correctly to (static) site characteristics, or because the available hydrometeorological variables at the sites do not completely describe the problem; and 3) meteorological forcing error, for example, incorrect precipitation amounts in the inputs at particular grid cells. The selected SNOTEL datasets were concentrated in the western United States and biased toward high-elevation sites. The precipitation was possibly underestimated by gridMet due to high heterogeneity within the grid cells and sparse data sampling. We attempted to decipher the error types while going through the results.

3. Results and discussion

In the following, we discuss the results from the forward model alone (section 3a). We then compare the DI models that integrate either SWE or SCF with different lag times to improve the prediction at monitored sites (section 3b) and investigate the sources of errors and factors controlling where DI is most useful (section 3c).

a. Forward model

In this work, the sequence-to-sequence LSTM without DI was employed as a valuable benchmark for comparison. The temporal test demonstrated that the LSTM could achieve promising results for the sites where SWE history was available. Across the 525 SNOTEL sites, the median NSE during the test period was 0.92, indicating that the model explained 92% of the observed variance. The forward model ($\mathrm{LSTM}_{\mathrm{temporal\_test}}$) yielded an RMSE of 48.2 mm, a bias of 5.3 mm, and a $d_{\max}$ of 62.4 mm (Table 1 and Fig. 1), which mainly reflect temporal and forcing errors. However, their relative importance is unknown. The forward model presented in this study demonstrates superior performance when benchmarked against the literature. For context, Garousi-Nejad and Tarboton (2022) evaluated SWE in the National Water Model, averaged SWE values across all SNOTEL sites, and obtained an NSE of 0.75 and a bias of −55 mm. Hill et al. (2019) obtained an RMSE of 59 mm by converting observed snow depths to SWE. GlobSnow v3.0 showed an RMSE of 71 mm for the whole winter (September–June) from 1979 to 2018 for North America (Venäläinen et al. 2021). Overall, the LSTM model developed in this study can be regarded as a state-of-the-art model for SWE prediction, as evidenced by its performance metrics.

The sites with higher model performance are scattered across the western United States, albeit with multiple spatial "pockets" exhibiting significant bias (Fig. 3a). In general, higher NSE values were found toward the northeastern part of the domain (with the exception of a few low-NSE sites on the boundary), and the lowest values were found toward the southwest. Large negative biases were present in California (Sierra Nevada range), western Oregon, Washington, and central Idaho, which could be attributed to pronounced and systematic errors in the gridded forcing dataset (see the detailed

TABLE 1. Model performances with different test scenarios. Temporal test means that the model was trained on some sites in a certain time period and tested on the same sites in a different period. Spatial test means that the model was trained on some sites and tested on other sites (in the same time period). In a spatiotemporal test, the model was trained on some sites in a period and tested on other sites in a different period. The time periods after the model name in the first column represent the training and testing periods, i.e., [training period]–[testing period]. DI(x, T) indicates a data integration LSTM model trained with either SWE or SCF as x that incorporated recent data via DI with a T-day lag.

Model	Bias (mm)	RMSE (mm)	R	NSE	d_{max} (mm)
Forward LSTM–temporal test [2001–15]–[2016–19]	5.3	48.2	0.97	0.92	62.4
Forward LSTM–temporal test with in situ forcing [2001–15]–[2016–19]	−0.01	39.3	0.98	0.94	50
Forward LSTM–spatial test [2001–15]–[2001–15]	3.2	53.9	0.96	0.88	70.9
Forward LSTM–spatial test with in situ forcing [2001–15]–[2001–15]	2.88	45.8	0.97	0.91	56.6
Forward LSTM–spatiotemporal test [2001–15]–[2016–19]	−1.2	65.8	0.95	0.84	93.8
DI(SWE, 30)–temporal test [2001–15]–[2016–19]	3.7	28.4	0.99	0.97	27.1
DI(SWE, 30)–spatial test [2001–15]–[2001–15]	2.0	32.6	0.98	0.96	29.1
DI(SWE, 30)–spatiotemporal test [2001–15]–[2016–19]	2.8	36.8	0.98	0.95	33.8
DI(SCF, 30)–temporal test [2001–15]–[2016–19]	−2.7	45.9	0.97	0.92	57.3
DI(SCF, 30)–spatial test [2001–15]–[2001–15]	0.6	54.3	0.96	0.87	67.9
DI(SCF, 30)–spatiotemporal test [2001–15]–[2016–19]	−3.27	66.8	0.95	0.84	86.6
DI(SWE, 7)–temporal test [2001–15]–[2016–19]	1.0	12.3	1.00	1.00	9.7
DI(SCF, 7)–temporal test [2001–15]–[2016–19]	−1.1	42.8	0.97	0.93	58.2

discussion in section 3c for Fig. 6) and/or higher interannual variability resulting from intermittent and extreme precipitation (e.g., atmospheric rivers; Goldenson et al. 2018). Overall, southern sites were more challenging to predict than northern ones, with low-latitude sites in Arizona (located at the southern edge of the domain) showing mixed results with a higher fraction of mid-to-low-performing sites. This pattern suggests the LSTM is more effective in capturing SWE in regions with climatologically deep snowpacks during the winter than ephemeral snow in the south (Hatchett 2021).

The spatial test, which can help to quantify spatial and forcing errors, revealed a noticeable declination of performance relative to the temporal test, which highlights the challenges facing spatial interpolation. The RMSE increased to 53.9 mm, and the NSE dropped to 0.88, with an R of 0.96. The NSE and R still seem competitive when compared with the results reported in the literature. For example, SNODAS, which regularly assimilates SWE observations and is considered a high-quality dataset, had an R of 0.93–0.98, while some blended satellite products (ATMS and AMSR2) had an R of approximately 0.9 (Gan et al. 2021). The maximum SWE error, d_{max}, was 70.9 mm for the spatial test, which was effectively the same as the 71-mm value calculated from the UA dataset. Additionally, the UA dataset uses all the data (and is thus not a spatial extrapolation or cross validation test where the testing sites are withheld from the input data), so this comparison is biased against the LSTM. However, it is worth noting that the gridded UA dataset represents specific attributes within a given area, whereas our LSTM is trained on point SWE measurements, which may be a closer representation of SNOTEL data than the gridded datasets. When the forward model was applied both for spatial and temporal applications, RMSE and d_{max} increased noticeably, by 17.6 and 31.4 mm, respectively, when compared with the temporal test (Table 1, Forward LSTM–spatiotemporal test).

Despite its high performance, the LSTM still encountered some challenges in predicting SWE because of the complex nature of snow processes—for example, long persistence, aerodynamic redistribution, and high spatial heterogeneity. SWE accumulation may be influenced by site-specific factors such as wind-driven redistribution patterns (Winstral and Marks 2002; Freudiger et al. 2017) or avalanches (Lehning and Fierz 2008) that could not be adequately described by available static attributes. Similar to forcing error, for snow the errors with precipitation do not dissipate but can accumulate over time, potentially causing errors for the entire snow season from the beginning of the accumulation phase. In the following analysis, the forward model was used to diagnose different types of errors and to serve as a benchmark for the DI models.

b. DI models integrating SWE or SCF

After establishing the forward LSTM model as a baseline, we investigated simulations that integrated either SWE or SCF data. When integrating 30-day lagged SWE, the model's performance improved to reach an NSE value of 0.97, RMSE of 28.4 mm, and d_{max} of 27.1 mm in the temporal test [Table 1; DI(SWE, 30)–temporal test]. Relative to the previous temporal test of the forward model, the inclusion of DI and 30-day-lagged SWE observations reduced the RMSE and d_{max} by 41% and 57%, respectively, while also greatly reducing bias to 3.7 mm. The NSE and d_{max} curves of the DI model in the temporal tests exhibited a significant contrast when compared with those of the forward model (Fig. 2; NSE and d_{max}). The contrast is more prominent than in some previous assimilation work based on ensemble Kalman filtering (Slater and Clark 2006). In a practical application, the DI SWE model will most likely be utilized to forecast SWE at monitored locations one month in advance. DI greatly reduced the bias observed in the northwest (Fig. 3a) and improved the model accuracy at low-NSE sites on the northeast boundary of the map caused by errors existing in the gridded forcing data or higher interannual variability (Fig. 3b). However, the model still had difficulty capturing the ephemeral snow at low-latitude sites in Arizona. The spatiotemporal test for DI(SWE, 30) did not use the test sites for training but only used

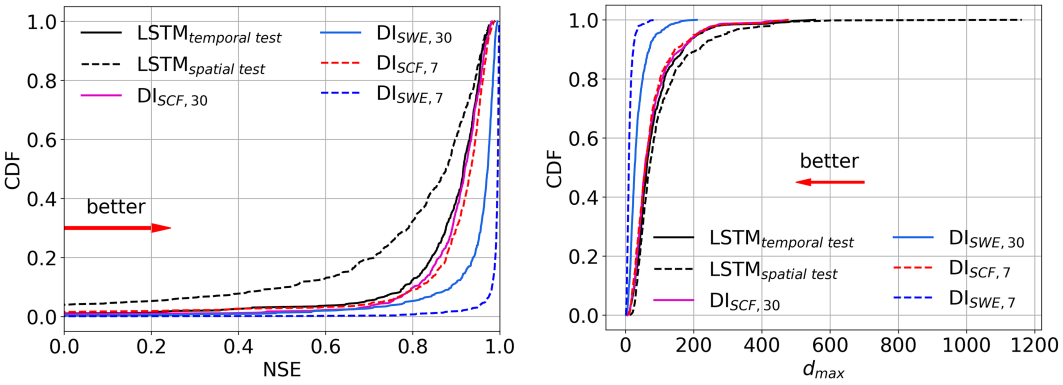


FIG. 2. Empirical cumulative distribution function of test performance metrics: (left) NSE and (right) d_{max} for the forward and DI models. $LSTM_{temporal\_test}$ (black solid line) denotes the forward model in the temporal test; $LSTM_{spatial\_test}$ (black dashed line) denotes the forward model in the spatial test (prediction at untrained sites). The remaining model results are for the temporal test: $DI_{SCF,30}$ (pink solid line) integrates 30-day-lagged snow cover fraction (SCF) observations; $DI_{SWE,30}$ (blue solid line) integrates 30-day-lagged snow water equivalent (SWE) observations; $DI_{SCF,7}$ (red dashed line) integrates 7-day-lagged SCF observations; and $DI_{SWE,7}$ (blue dashed line) integrates 7-day-lagged SWE observations. The red arrows marked "better" show the direction in which model improvement occurs.

the 30-day lagged SWE observations for DI in testing. The errors for the spatiotemporal test rose modestly (in absolute terms) in RMSE (36.8 mm) and d_{max} (33.8 mm), indicating that this model could be useful for newly instrumented SWE observational sites or campaigns without long data records for training. It is important to note that this prediction used reanalysis meteorological forcing data, and thus did not seek to represent internal variability or uncertainty caused by the meteorological forecast data.

Integrating 30-day-lagged SCF only had minor impacts on d_{max} and NSE in the temporal test (red lines in Fig. 2) and was more effective in reducing temporal errors rather than spatial errors. A comparison between the forward model and DI(SCF, 30) in the temporal test showed a reduction of 2.3 and 5.1 mm in RMSE and d_{max}, respectively. However, in the spatial test, d_{max} was reduced by 3 mm but RMSE increased by 0.4 mm (Table 1). Therefore, SCF's most effective role was to reduce temporal errors.

While for all the sites the benefit of incorporating SCF data did not seem significant, it improved SWE prediction for sites with shallow snow depth, especially during the snowmelt period. Site stratification shows that SCF brings benefits to shallow-snow sites (Fig. 4). Table A3 in the appendix provides the metrics of DI(SCF, 30) and the forward model at sites where mean SWE is less than 40 mm. Both the forward model and DI(SCF, 30) have better performance during snow accumulation rather than melt, as snow ablation is a more complicated process. However, integrating 30-day-lagged SCF greatly decreased bias and improved the NSE for SWE prediction during the melt season at shallow-snow sites. This is presumably because SCF has a more linear relationship with SWE during the melt season (Swenson and Lawrence 2012). Since the snowmelt phase has always been a challenge, we think SCF is still useful for predicting available snowmelt water at large scales.

DI(SWE, 7) prediction reached nearly perfect simulations in the temporal test with R and NSE both close to 1.00, a nearly negligible bias of 1.0 mm, RMSE of 12.3 mm, and a d_{max} of 9.7 mm. The remaining RMSE could be explained by the effect of 7-day differences in actual and input precipitation. Previously, Leisenring and Moradkhani (2011) employed multiple data assimilation methods to assimilate SWE data "whenever new observations are available" (which would be similar to a 1-day lag) for a Ward Creek SNOTEL site in California. They reported that EnKF produced an RMSE of 47.24 mm and an NSE of 0.936, while another method called EPF-WRR produced an RMSE of 28.95 mm and an NSE of 0.976. In contrast, our DI(SWE, 7) yielded an NSE of 0.995 for the same site. It seems that deep network-based data integration is highly performant and well suited for modeling the snow processes during the 7 days before prediction.

Upon examination of several sites, it was observed that DI often corrected the underestimation of snow accumulation by the forward models (Fig. 5a). The DI and forward models diverged early (about a month) into the snow accumulation season. It was surprising to find that, even with 30-day lagged SWE, the DI(SWE, 30) model was closer to the observed SWE than the forward only model. This stands in contrast to the general trend observed in other hydrological LSTM models, where only the most recent observations typically contribute to predictive improvements (Feng et al. 2020). DI(SWE, 30) consistently followed the observed snow accumulation trend from the beginning of all accumulation seasons (Figs. 5a,c). However, there were cases where DI(SWE, 30) and DI(SWE, 7) diverged earlier, for example, December 2016, October 2017, and December 2018 (Fig. 5d), reflecting the impact of ephemeral snow accumulation and melt. Models integrating SCF were found to be intermediate in NSE but significantly underestimated peak SWE for deep snow sites (Fig. 5a). Nevertheless, SCF still had benefits for predicting ephemeral snow at the beginning of the accumulation season and the end of the melt season, and was able to correct the errors in some testing periods of DI(SWE, 30) and the forward model, e.g., from March to May 2017 and 2019 of Fig. 5b and from October to November 2017 of Fig. 5d.

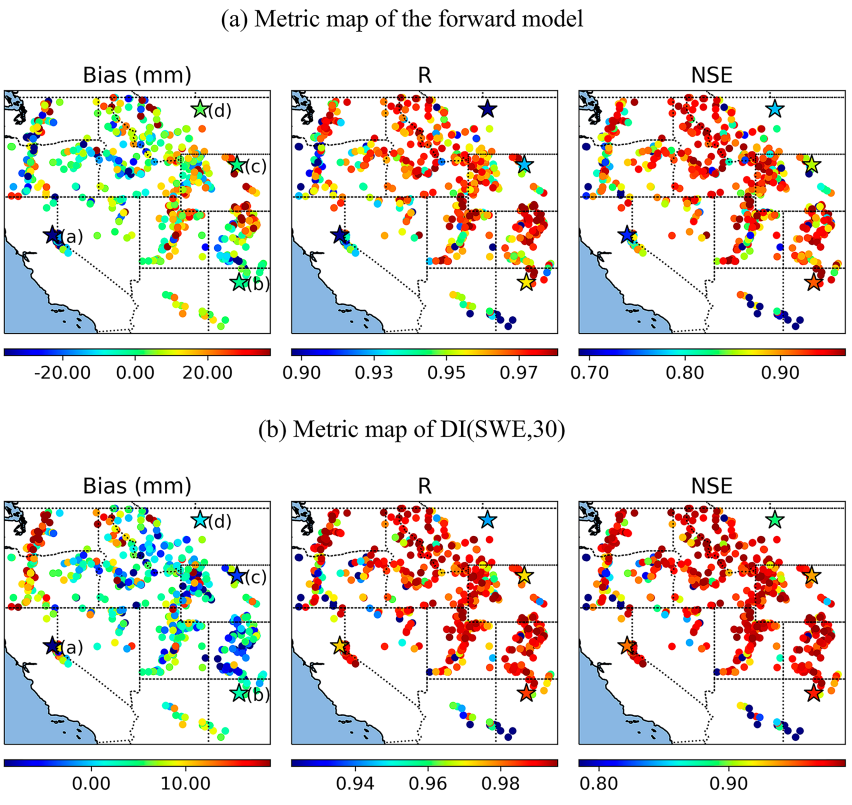


FIG. 3. Map of temporal test performance metrics (bias, R, and NSE) for (a) the forward model and (b) the DI model integrating SWE observations with a 30-day lag, DI(SWE, 30). The sites annotated in the map, represented by star-shaped points and labeled with letters a–d, represent the locations for the time series plots shown in Fig. 5, described below.

A question of interest is, *How could the LSTM be so effective at integrating observations of varied lags*? We hypothesize that LSTM may simply build a model that accounts for the differences between SWE^{t-T} and SWE^{t}. In other words, it only needs to learn to represent the snow processes in the period of T by giving an accurate initial condition at day $t - T$. LSTM could also build longer-term memory to internally keep track of states related to thermal and compaction processes. In addition, LSTM could potentially have units that perform uncertainty analyses, as we showed previously (Fang et al. 2020). This would enable LSTM to weigh the importance of the information based on its perceived uncertainty. Because of LSTM's strong DI capability, DI(SWE, T) essentially serves as an *error stopper* where errors longer than T are suppressed. We leverage LSTM's adaptive DI ability as an error detector in the next section: we show sites with long-term (multimonth), short-term (7–30 days), and ultra-short-term (<7 days) forcing errors and the approximate magnitudes of these different error sources.

c. Error source analysis

In this section we extract a few insights from model comparison. Recall that we previously considered three types of errors, due to (type A: temporal error) model temporal dynamics or nonstationarity, (type B: spatial error) spatial heterogeneity, and (type C: forcing error) inaccurate forcings. The models' errors in tests can be a combination of different types of errors. However, these three types of errors are most likely not additive—meaning the errors from different sources cannot be simply added to estimate the total error (neither the unit of the error metrics, e.g., RMSE, or the nature of the errors is linear). Nonetheless, separating them roughly can be helpful to diagnose the magnitude of the error sources and to anticipate the likely effectiveness of improvement strategies.

While gridMet was employed as the main forcing data to ensure the model's wide applicability, we contrasted this forcing with precipitation and temperature recorded at SNOTEL sites (missing in situ temperature data were substituted with gridMet temperature), which shows gridded forcing data are an important (but not the sole) source of error. The forward model, when utilizing SNOTEL in situ forcing, exhibited significant improvement, leading to a reduction of 8.9 and 8.1 mm in RMSE and a reduction of 12.4 and 14.3 mm in $d_{\max}$ (Table 1) in the temporal and spatial tests, respectively. In situ forcing data decreased the RMSE at sites with large bias in western Oregon and Washington, as well as the northeast domain (Figs. 6a,b), but did not fundamentally change the behavior of the models at sites

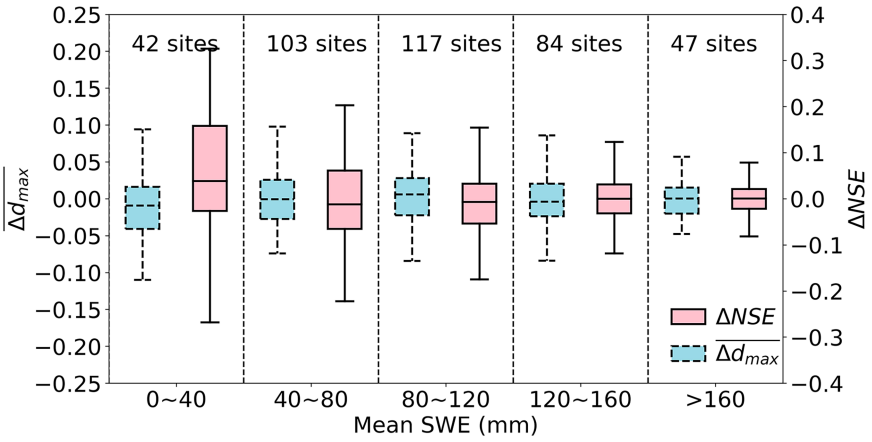


FIG. 4. Boxplots for $\overline{\Delta d_{\max}}$ (blue boxes outlined with dashed lines) and ΔNSE (pink boxes outlined with solid lines) stratified by mean SWE; $\Delta d_{\max}$ and ΔNSE are the differences of $d_{\max}$ and NSE, respectively, between DI(SCF, 30) and the forward model in the spatial tests. The $\overline{\Delta d_{\max}}$ is $\Delta d_{\max}$ normalized by maximum SWE. The negative $\Delta d_{\max}$ and positive ΔNSE values denote that the DI(SCF, 30) model has improved performance over the forward model.

in California near Lake Tahoe (location a, Fig. 5a, yellow line). Thus, the errors for the Californian sites were not mainly caused by issues with gridMet forcing data. Although the reason is still under investigation, a plausible explanation is that we need a longer training period (>15 years) for these sites due to the high interannual variability of precipitation events (e.g., atmospheric rivers) that drive snowfall totals (and SWE) in the Sierra Nevada range. Moreover, local aerodynamic redistribution

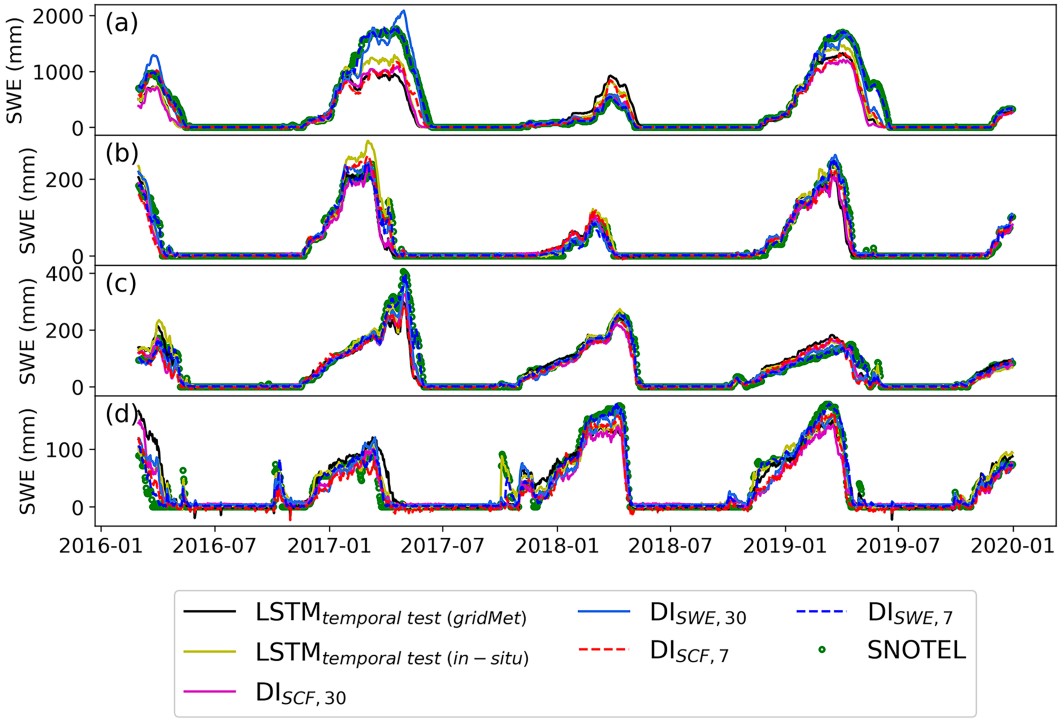
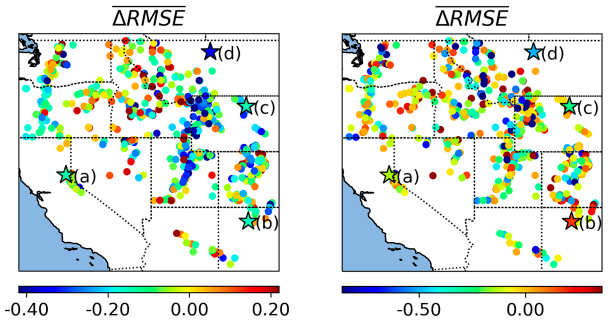


FIG. 5. Time series of the forward model ($LSTM_{temporal\_test}$ with either gridMet or in situ forcings) and the models integrating SWE and SCF at different time lags for four sites, shown for the temporal tests. The locations of these sites are annotated in Fig. 3 and in Figs. 6, and 7, described below. Elevations are (a) 2101, (b) 2621, (c) 2548, and (d) 1423 m. These sites were selected because of their large discrepancies between the forward and DI models.



(a) Temporal test (b) Spatial test

FIG. 6. Maps of RMSE differences between forward models with and without in situ precipitation and temperature for the (a) temporal test and (b) spatial test. $\overline{\Delta\text{RMSE}}$ is ΔRMSE normalized by $d_{\max}$ of the forward model using gridMet. This figure shows the fraction of accumulated SWE error that could be reduced by using in situ forcing data. A large negative value means that using in situ forcing was highly effective, and a positive value means in situ data actually increased the error. Overall, in situ forcing leads to better models than the gridded forcing. The sites annotated in the map, represented using star-shaped points and labeled with letters a–d, represent the locations for the time series plots in Fig. 5.

patterns may play an important, systematic role for these locations, leading to errors not capturable by this model.

Subsequently, we can take the spatiotemporal test of the forward model (RMSE = 65.8 mm and $d_{\max}$ = 93.8 mm; Table 1) as a worst-case scenario with all prediction errors represented. Its difference from the forward model's spatial test (ΔRMSE = 65.8 − 53.9 = 11.9 mm; $\Delta d_{\max}$ = 93.8 − 70.9 = 22.9 mm) is likely due to model errors with respect to multiyear nonstationarity (type A, temporal error). Relatedly, its difference from the forward model's temporal test (ΔRMSE = 65.8 − 48.2 = 17.6 mm; $\Delta d_{\max}$ = 93.8 − 62.4 = 31.4 mm) is connected to unexplained spatial heterogeneity (type B, spatial error). The sites susceptible to spatial error (Fig. 7a) concentrate on the northeastern part of the domain, the northwestern mountain ranges, and Arizona. These sites may have some (currently unclear) distinct characteristics that are difficult to learn based on the current input attributes. Assuming the above, then the difference between the temporal tests of the forward model and DI(SWE, 30) (ΔRMSE = 48.2 − 28.4 = 19.8 mm; $\Delta d_{\max}$ = 62.4 − 27.1 = 35.3 mm) is related to both the impact of model dynamics (type A) and the cumulative effect of forcing errors (type C) between 30 days and snowmelt. Such errors are carried for a long time and, if not reduced with DI, persist longer to cause large errors for the whole season. In addition, the sites with large ΔNSE here are not the ones with large spatial errors, except for those in Arizona (Fig. 7b), indicating that they are caused by different processes, such as the complicated temporal dynamics caused by high interannual variability of precipitation events at location a. Relatedly, the difference between the temporal tests of DI(SWE, 7) and DI(SWE, 30) (ΔRMSE = 28.4 − 12.3 = 16.1 mm; $\Delta d_{\max}$ = 27.1 − 9.7 = 17.4 mm) (see the map in Fig. 7c) is the error accumulated between 7 and 30 days. These sites (Fig. 7c) are different from the hotspots in Fig. 7b and should represent errors that are accumulated and then removed in a shorter time frame (7–30 days). The remaining error (RMSE = 12.3 mm; $d_{\max}$ = 9.7 mm) is accumulated for only 7 days. Again, the errors are not additive and are only intended for a rough analysis.

Sites with large ΔNSE between DI(SWE, 7) and DI(SWE, 30) also have substantial ephemeral snow, which can be susceptible to large forcing errors with the forward model. Examining the differences between the DI and forward models, we notice DI(SWE, 30) strongly improved NSE in parts of the domain but had little impact on some other sites (Fig. 7b). In general, ΔNSE tended to be higher in the south of the domain rather than the north, and was especially large along the southern boundary. However, the 30-day DI model still encountered some difficulties on low-latitude sites in the mountains of Arizona and New Mexico (Fig. 7c). A plausible explanation is that DI was not very helpful with the ephemeral snow situations occurring in those sites. The input data could miss precipitation events and temperature changes that led to errors—in October 2017 in Fig. 5d, for example, the forward model did not have a response at all but the model with in situ precipitation well captured the observation trends. By the time DI(SWE, 30) encountered this discrepancy, the ephemeral snow would have already melted.

4. Conclusions

The LSTM forward model was already highly competitive in comparison with previous SWE models, and data integration with LSTM was effective at increasing prediction performance. Integrating SWE at a 30-day lag, the median NSE was as high as 0.97 for monitored SNOTEL sites, with a negligible bias and a small $d_{\max}$ of 27.1 mm. Such effectiveness is because snow modeling errors have long time persistence, which DI is effective at addressing. This means such a model could serve as a useful forecast scheme for water resource managers in snow-dominated regions. That being said, there are sites with significant numbers of ephemeral snow events where 1-month-ahead predictions would not have much benefit, and a shorter lag would be needed. It is much more difficult to predict at monthly scale and high accuracy for these sites with ephemeral snow.

Using LSTM and DI as probes for error sources, our analysis showed that both the temporal and spatial errors are important. Temporal errors can be effectively reduced by DI with SWE, while spatial errors are more difficult to reduce—assimilating SCF data mainly had benefits only for shallow-snow sites, and especially during snowmelt. We need to seek more extensive spatial observations like snow depths to constrain the simulations, which will be particularly useful for the hotspots identified in Fig. 7b. Further, our model comparisons show that hotspots of different kinds of errors exist in different parts of the domain. The highlighted spatial error exists in the northern and southern boundaries of the domain; the majority of the multimonth error occurs in California and Oregon, and on the northeastern and southern boundaries of

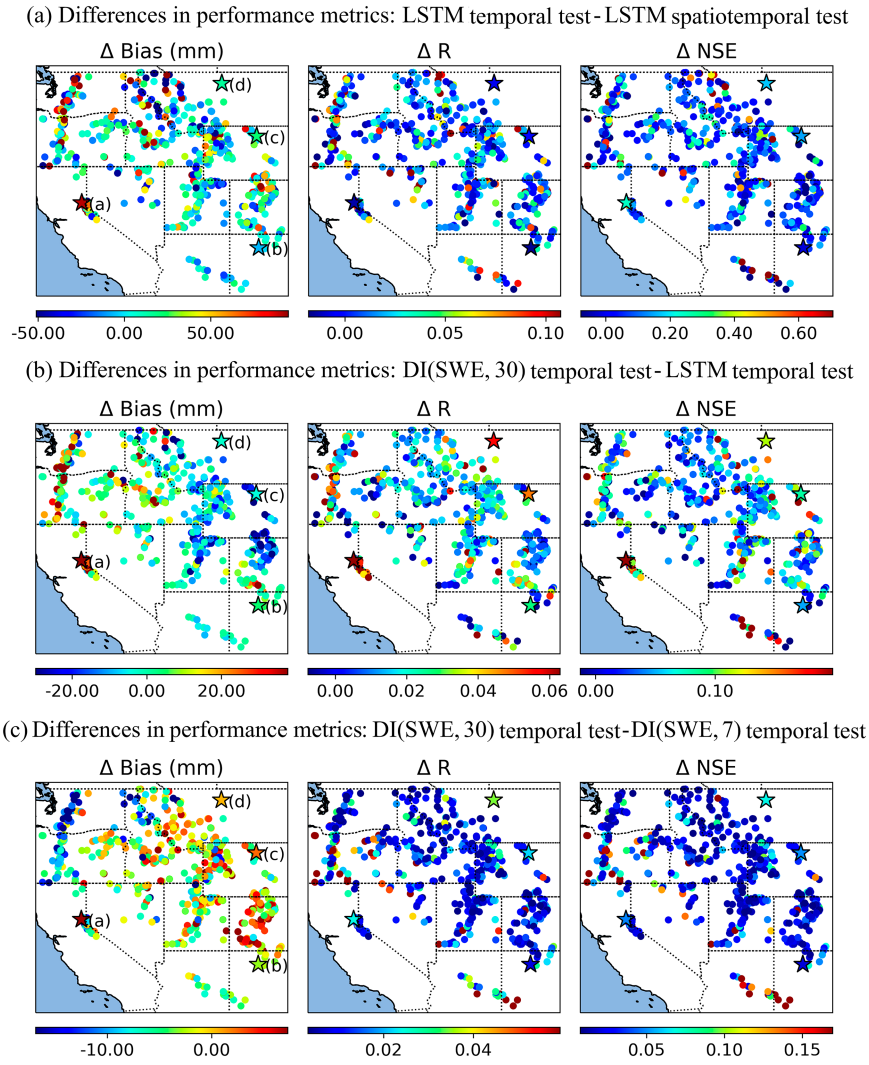


FIG. 7. Maps of performance metric (bias, *R*, and NSE) differences between models: (a) metric differences between the temporal and spatiotemporal tests of the forward model, (b) metric differences between the temporal tests of the data integration model incorporating SWE observations with a 30-day lag, DI(SWE, 30), and the forward model, and (c) metric differences between the temporal tests of DI(SWE, 30) and DI(SWE, 7) (7-day lag). The sites annotated in the map, represented using star-shaped points and labeled with letters a–d, represent the locations for the time series plots in Fig. 5.

our domain; and the error of less than 1 month concentrates in more southern mountains and where ephemeral snow occurs frequently. For climate change impact simulations, we cannot rely on DI, and would require better resolution of precipitation processes and longer data history in the climate models to improve SWE prediction in these regions (Rhoades et al. 2018). Considering the power of LSTM, however, it is unlikely that improved structures of process-based snow models will obtain noticeably lower errors than LSTM's forward model.

In real applications, DI in a SWE model can be beneficial for filling data gaps at sites that lack daily data, which is especially pertinent when investigating long-term trends and variations in snowpack. For instance, in the eastern United States where SWE is documented on a weekly or biweekly basis, DI can be a crucial tool in overcoming data insufficiency. Once accurate meteorological forecast data are available, the DI model can be employed to forecast SWE at monitored sites approximately one month in advance. The results presented in this work have set up benchmark levels and provide guidance for future improvements.

Acknowledgments. This work was supported by the Office of Biological and Environmental Research of the U.S. Department of Energy under Contract DE-SC0016605. Computing was partially supported by U.S. National Science Foundation Award PHY 2018280. The National Center for

Atmospheric Research is sponsored by the National Science Foundation.

Data availability statement. The deep learning code relevant to this work (http://doi.org/10.5281/zenodo.5015120), SNOTEL data (https://www.nrcs.usda.gov/wps/portal/wcc/home/), MODIS SCF data (https://nsidc.org/data/mod10a1f/versions/61), and GridMet forcing data (https://www.climatologylab.org/gridmet.html) are all available online.

APPENDIX

Sensitivity Experiments

We ran many experiments, including varying LSTM sequence length and static attributes. This was done not to choose which model to show, but rather to demonstrate the impact of certain configurations on the results to provide some insights.

We explored the effects of sequence length on LSTM performance. Table A1 provides a performance comparison of the forward model, DI(SWE, 30), and DI(SWE, 7) with different sequence lengths (rho = 90 and 365) in the temporal tests (all sites were trained in 2001–15 and tested in 2016–19). Increasing the sequence length to 365 can benefit the forward model, considering the period of snow dynamics is one year for seasonal and ephemeral snowpacks characteristic of the western United States. However, the DI models only need a short sequence length.

Although not used for model selection, we tested the following static attributes: (i) latitudes; (ii) soil depth; (iii) topographic characteristics, including elevation, slope, and aspect; (iv) land cover characteristics (dominant land cover, dominant land cover fraction, forest fraction) and vegetation characteristics (rooting depth) from MODIS Land Cover Type Yearly L3 Global 500 m (MCD12Q1); and (v) geological characteristics (subsurface porosity and permeability) from the GLobal HYdrogeology MaPS (GLHYMPS) datasets. We ran spatial tests with different combinations of static attributes to show the impacts of adding or removing attributes (Table A2). It turned out that latitude, topographic, and land cover characteristics all had positive impacts on the spatial test results, but, as expected, the geological characteristics and soil depth had no impact.

We further compared the performance of the forward model, DI(SWE, 30), and DI(SCF, 30) in the snow accumulation and melt seasons in the spatial tests (Table A3). All models have better performance in the accumulation season than the melt season since snow ablation is more complicated than accumulation. More interesting, for the shallow-snow sites where mean SWE was less than 40 mm, the integration of 30-day-lagged SCF significantly improved the performance in terms of NSE and bias because of a more linear relationship between SCF and SWE during snowmelt (Swenson and Lawrence 2012).

TABLE A1. Temporal test results for the forward model, DI(SWE, 30), and DI(SWE, 7) with different values for rho (sequence length: 90 and 365).

Model	Bias (mm)	RMSE (mm)	R	NSE	d_{max} (mm)
Forward–rho = 90 [2001–15]–[2016–19]	−18.49	57.8	0.96	0.88	88.5
Forward–rho = 365 [2001–15]–[2016–19]	5.3	48.2	0.97	0.92	62.4
DI(SWE, 30)–rho = 90 [2001–15]–[2016–19]	−2.8	27.6	0.99	0.97	24.2
DI(SWE, 30)–rho = 365 [2001–15]–[2016–19]	3.7	28.4	0.99	0.97	27.1
DI(SWE, 7)–rho = 90 [2001–15]–[2016–19]	−0.7	11.5	1.00	1.00	8.8
DI(SWE, 7)–rho = 365 [2001–15]–[2016–19]	1.0	12.3	1.00	1.00	9.7

TABLE A2. Spatial test results from models with different combinations of the static variables. For each model, we ran 10-fold cross validation where the model was trained on a rotating group of the monitored sites (9/10 site groups) from 2001 to 2015 and tested on the remaining untrained sites from 2001 to 2015. The model performance listed as the spatial test result is the median of the 10 rounds of testing.

Model	Bias (mm)	RMSE (mm)	R	NSE	d_{max} (mm)
Forward—with latitudes, topographic, and land cover, vegetation, and geological characteristics [2001–15]–[2001–15]	−0.05	59	0.96	0.86	78.4
Forward—with latitudes, topographic, and land cover and vegetation characteristics [2001–15]–[2001–15]	4.9	59.5	0.96	0.87	72.7
Forward—with latitudes, topographic, and land cover characteristics [2001–15]–[2001–15]	3.17	53.9	0.96	0.88	70.9
Forward—with topographic and land cover characteristics [2001–15]–[2001–15]	5.6	58.2	0.96	0.87	75.0
Forward—with latitudes and topographic [2001–15]–[2001–15]	4.5	54.7	0.96	0.89	72.9
Forward—with latitudes [2001–15]–[2001–15]	11.8	67.6	0.95	0.83	90.0
Forward—without any static attributes [2001–15]–[2001–15]	7.0	65.9	0.95	0.84	85.0

TABLE A3. Spatial test results for the forward model, DI(SWE, 30), and DI(SCF, 30) in the snow accumulation and melt seasons. The last four rows are the spatial test results from the forward model and DI(SCF, 30) at shallow-snow sites where mean SWE was less than 40 mm.

Model	Bias (mm)	RMSE (mm)	*R*	NSE
Forward—snow accumulation [2001–15]–[2001–15]	2.4	47.9	0.98	0.9
Forward—snowmelt [2001–15]–[2001–15]	3.1	58.0	0.95	0.82
DI(SWE, 30)—snow accumulation [2001–15]–[2001–15]	2.1	24.7	0.99	0.98
DI(SWE, 30)—snowmelt [2001–15]–[2001–15]	2.3	37.7	0.98	0.94
DI(SCF, 30)—snow accumulation [2001–15]–[2001–15]	−0.2	47.9	0.98	0.9
DI(SCF, 30)—snowmelt [2001–15]–[2001–15]	−1.6	61.4	0.95	0.82
Forward—snow accumulation (42 sites with mean SWE < 40 mm) [2001–15]–[2001–15]	14.3	36.1	0.91	0.66
Forward—snowmelt (42 sites with mean SWE < 40 mm) [2001–15]–[2001–15]	9.0	21.4	0.87	0.1
DI(SCF, 30)—snow accumulation (42 sites with mean SWE < 40 mm) [2001–15]–[2001–15]	6.4	34.9	0.91	0.65
DI(SCF, 30)—snowmelt (42 sites with mean SWE < 40 mm) [2001–15]–[2001–15]	1.8	19.1	0.86	0.24

REFERENCES

Abatzoglou, J. T., 2013: Development of gridded surface meteorological data for ecological applications and modelling. *Int. J. Climatol.*, **33**, 121–131, https://doi.org/10.1002/joc.3413.

Aboelyazeed, D., C. Xu, F. M. Hoffman, J. Liu, A. W. Jones, C. Rackauckas, K. Lawson, and C. Shen, 2023: A differentiable, physics-informed ecosystem modeling and learning framework for large-scale inverse problems: Demonstration with photosynthesis simulations. *Biogeosciences*, **20**, 2671–2692, https://doi.org/10.5194/bg-20-2671-2023.

Afzaal, H., A. A. Farooque, F. Abbas, B. Acharya, and T. Esau, 2020: Groundwater estimation from major physical hydrology components using artificial neural networks and deep learning. *Water*, **12**, 5, https://doi.org/10.3390/w12010005.

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

Clark, M. P., D. E. Rupp, R. A. Woods, X. Zheng, R. P. Ibbitt, A. G. Slater, J. Schmidt, and M. J. Uddstrom, 2008: Hydrological data assimilation with the ensemble Kalman filter: Use of streamflow observations to update states in a distributed hydrological model. *Adv. Water Resour.*, **31**, 1309–1324, https://doi.org/10.1016/j.advwatres.2008.06.005.

Corriveau, J., P. A. Chambers, A. G. Yates, and J. M. Culp, 2011: Snowmelt and its role in the hydrologic and nutrient budgets of prairie streams. *Water Sci. Technol.*, **64**, 1590–1596, https://doi.org/10.2166/wst.2011.676.

——, ——, and J. M. Culp, 2013: Seasonal variation in nutrient export along streams in the northern Great Plains. *Water Air Soil Pollut.*, **224**, 1594, https://doi.org/10.1007/s11270-013-1594-1.

Dawson, N., P. Broxton, and X. Zeng, 2017: A new snow density parameterization for land data initialization. *J. Hydrometeor.*, **18**, 197–207, https://doi.org/10.1175/JHM-D-16-0166.1.

——, ——, and ——, 2018: Evaluation of remotely sensed snow water equivalent and snow cover extent over the contiguous United States. *J. Hydrometeor.*, **19**, 1777–1791, https://doi.org/10.1175/JHM-D-18-0007.1.

Diro, G. T., and H. Lin, 2020: Subseasonal forecast skill of snow water equivalent and its link with temperature in selected SubX models. *Wea. Forecasting*, **35**, 273–284, https://doi.org/10.1175/WAF-D-19-0074.1.

Egli, L., and T. Jonas, 2009: Hysteretic dynamics of seasonal snow depth distribution in the Swiss Alps. *Geophys. Res. Lett.*, **36**, L02501, https://doi.org/10.1029/2008GL035545.

Evensen, G., 2009: Ensemble methods. *Data Assimilation: The Ensemble Kalman Filter*, G. Evensen, Ed., Springer, 119–137, https://doi.org/10.1007/978-3-642-03711-5_9.

Fang, K., and C. Shen, 2020: Near-real-time forecast of satellite-based soil moisture using long short-term memory with an adaptive data integration kernel. *J. Hydrometeor.*, **21**, 399–413, https://doi.org/10.1175/JHM-D-19-0169.1.

——, ——, D. Kifer, and X. Yang, 2017: Prolongation of SMAP to spatiotemporally seamless coverage of continental U.S. using a deep learning neural network. *Geophys. Res. Lett.*, **44**, 11 030–11 039, https://doi.org/10.1002/2017GL075619.

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

——, ——, ——, D. Feng, and C. Shen, 2022: The data synergy effects of time-series deep learning models in hydrology. *Water Resour. Res.*, **58**, e2021WR029583, https://doi.org/10.1029/2021WR029583.

Feng, D., K. Fang, and C. Shen, 2020: Enhancing streamflow forecast and extracting insights using long-short term memory networks with data integration at continental scales. *Water Resour. Res.*, **56**, e2019WR026793, https://doi.org/10.1029/2019WR026793.

——, K. Lawson, and C. Shen, 2021: Mitigating prediction error of deep learning streamflow models in large data-sparse regions with ensemble modeling and soft data. *Geophys. Res. Lett.*, **48**, e2021GL092999, https://doi.org/10.1029/2021GL092999.

——, J. Liu, K. Lawson, and C. Shen, 2022: Differentiable, learnable, regionalized process-based models with multiphysical outputs can approach state-of-the-art hydrologic prediction accuracy. *Water Resour. Res.*, **58**, e2022WR032404, https://doi.org/10.1029/2022WR032404.

Freudiger, D., I. Kohn, J. Seibert, K. Stahl, and M. Weiler, 2017: Snow redistribution for the hydrological modeling of Alpine catchments. *Wiley Interdiscip. Rev.: Water*, **4**, e1232, https://doi.org/10.1002/wat2.1232.

Gan, Y., Y. Zhang, C. Kongoli, C. Grassotti, Y. Liu, Y.-K. Lee, and D.-J. Seo, 2021: Evaluation and blending of ATMS and AMSR2 snow water equivalent retrievals over the

conterminous United States. *Remote Sens. Environ.*, **254**, 112280, https://doi.org/10.1016/j.rse.2020.112280.

Garousi-Nejad, I., and D. G. Tarboton, 2022: A comparison of National Water Model retrospective analysis snow outputs at snow telemetry sites across the western United States. *Hydrol. Processes*, **36**, e14469, https://doi.org/10.1002/hyp.14469.

Girotto, M., K. N. Musselman, and R. L. H. Essery, 2020: Data assimilation improves estimates of climate-sensitive seasonal snow. *Curr. Climate Change Rep.*, **6**, 81–94, https://doi.org/10.1007/s40641-020-00159-7.

Goldenson, N., L. R. Leung, C. M. Bitz, and E. Blanchard-Wrigglesworth, 2018: Influence of atmospheric rivers on mountain snowpack in the western United States. *J. Climate*, **31**, 9921–9940, https://doi.org/10.1175/JCLI-D-18-0268.1.

Graves, A., 2012: Long short-term memory. *Supervised Sequence Labelling with Recurrent Neural Networks*, A. Graves, Ed., Studies in Computational Intelligence, Vol. 385, Springer, 37–45, https://doi.org/10.1007/978-3-642-24797-2_4.

Hall, D. K., and G. A. Riggs, 2021: MODIS/Terra Snow Cover Daily L3 Global 500m SIN Grid, version 61. NSDIC, accessed 4 February 2022, https://doi.org/10.5067/MODIS/MOD10A1.061.

Hatchett, B. J., 2021: Seasonal and ephemeral snowpacks of the conterminous United States. *Hydrology*, **8**, 32, https://doi.org/10.3390/hydrology8010032.

——, A. M. Rhoades, and D. J. McEvoy, 2022: Monitoring the daily evolution and extent of snow drought. *Nat. Hazards Earth Syst. Sci.*, **22**, 869–890, https://doi.org/10.5194/nhess-22-869-2022.

Hill, D. F., E. A. Burakowski, R. L. Crumley, J. Keon, J. M. Hu, A. A. Arendt, K. Wikstrom Jones, and G. J. Wolken, 2019: Converting snow depth to snow water equivalent using climatological variables. *Cryosphere*, **13**, 1767–1784, https://doi.org/10.5194/tc-13-1767-2019.

Hochreiter, S., and J. Schmidhuber, 1997: Long short-term memory. *Neural Comput.*, **9**, 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735.

——, Y. Bengio, P. Frasconi, and J. Schmidhuber, 2001: Gradient flow in recurrent nets: The difficulty of learning long-term dependencies. *A Field Guide to Dynamical Recurrent Neural Networks*, S. C. Kremer and J. F. Kolen, Eds., IEEE Press, 237–244.

Houser, P. R., W. J. Shuttleworth, J. S. Famiglietti, H. V. Gupta, K. H. Syed, and D. C. Goodrich, 1998: Integration of soil moisture remote sensing and hydrologic modeling using data assimilation. *Water Resour. Res.*, **34**, 3405–3420, https://doi.org/10.1029/1998WR900001.

King, F., A. R. Erler, S. K. Frey, and C. G. Fletcher, 2020: Application of machine learning techniques for regional bias correction of snow water equivalent estimates in Ontario, Canada. *Hydrol. Earth Syst. Sci.*, **24**, 4887–4902, https://doi.org/10.5194/hess-24-4887-2020.

Klotz, D., F. Kratzert, M. Gauch, A. Keefe Sampson, J. Brandstetter, G. Klambauer, S. Hochreiter, and G. Nearing, 2022: Uncertainty estimation with deep learning for rainfall-runoff modelling. *Hydrol. Earth Syst. Sci.*, **26**, 1673–1693, https://doi.org/10.5194/hess-26-1673-2022.

Kopytkovskiy, M., M. Geza, and J. E. McCray, 2015: Climate-change impacts on water resources and hydropower potential in the upper Colorado River basin. *J. Hydrol.*, **3**, 473–493, https://doi.org/10.1016/j.ejrh.2015.02.014.

Koster, R. D., R. H. Reichle, and S. P. P. Mahanama, 2017: A data-driven approach for daily real-time estimates and forecasts of near-surface soil moisture. *J. Hydrometeor.*, **18**, 837–843, https://doi.org/10.1175/JHM-D-16-0285.1.

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

Lehning, M., and C. Fierz, 2008: Assessment of snow transport in avalanche terrain. *Cold Reg. Sci. Technol.*, **51**, 240–252, https://doi.org/10.1016/j.coldregions.2007.05.012.

Leisenring, M., and H. Moradkhani, 2011: Snow water equivalent prediction using Bayesian data assimilation methods. *Stochastic Environ. Res. Risk Assess.*, **25**, 253–270, https://doi.org/10.1007/s00477-010-0445-5.

Li, D., M. L. Wrzesien, M. Durand, J. Adam, and D. P. Lettenmaier, 2017: How much runoff originates as snow in the western United States, and how will that change in the future? *Geophys. Res. Lett.*, **44**, 6163–6172, https://doi.org/10.1002/2017GL073551.

Li, Q., Z. Wang, W. Shangguan, L. Li, Y. Yao, and F. Yu, 2021: Improved daily SMAP satellite soil moisture prediction over China using deep learning model with transfer learning. *J. Hydrol.*, **600**, 126698, https://doi.org/10.1016/j.jhydrol.2021.126698.

Luce, C. H., and D. G. Tarboton, 2004: The application of depletion curves for parameterization of subgrid variability of snow. *Hydrol. Processes*, **18**, 1409–1422, https://doi.org/10.1002/hyp.1420.

Ma, K., D. Feng, K. Lawson, W.-P. Tsai, C. Liang, X. Huang, A. Sharma, and C. Shen, 2021: Transferring hydrologic data across continents—Leveraging data-rich regions to improve hydrologic prediction in data-sparse regions. *Water Resour. Res.*, **57**, e2020WR028600, https://doi.org/10.1029/2020WR028600.

Magand, C., A. Ducharne, N. L. Moine, and S. Gascoin, 2014: Introducing hysteresis in snow depletion curves to improve the water budget of a land surface model in an Alpine catchment. *J. Hydrometeor.*, **15**, 631–649, https://doi.org/10.1175/JHM-D-13-091.1.

Magnusson, J., G. Nævdal, F. Matt, J. F. Burkhart, and A. Winstral, 2020: Improving hydropower inflow forecasts by assimilating snow data. *Hydrol. Res.*, **51**, 226–237, https://doi.org/10.2166/nh.2020.025.

McCuen, R. H., Z. Knight, and A. G. Cutter, 2006: Evaluation of the Nash–Sutcliffe Efficiency Index. *J. Hydrol. Eng.*, **11**, 597–602, https://doi.org/10.1061/(ASCE)1084-0699(2006)11:6(597).

Meyal, A. Y., R. Versteeg, E. Alper, D. Johnson, A. Rodzianko, M. Franklin, and H. Wainwright, 2020: Automated cloud based long short-term memory neural network based SWE prediction. *Front. Water*, **2**, 574917, https://doi.org/10.3389/frwa.2020.574917.

Nearing, G. S., and Coauthors, 2022: Technical note: Data assimilation and autoregression for using near-real-time streamflow observations in long short-term memory networks. *Hydrol. Earth Syst. Sci.*, **26**, 5493–5513, https://doi.org/10.5194/hess-26-5493-2022.

NOHRSC, 2004: Snow Data Assimilation System (SNODAS) data products at NSIDC, version 1. NSIDC, https://doi.org/10.7265/N5TB14TC.

Nowak, K., L. Bearup, D. Larsen, D. Garcia, C. Moore, and S. Baker, 2021: Emerging technologies in snow monitoring. U.S. Department of the Interior Rep. 508, 67 pp., https://www.usbr.gov/research/docs/news/Emerging_Snow_Monitoring_Report_508.pdf.

O, S., and R. Orth, 2021: Global soil moisture data derived through machine learning trained with in-situ measurements. *Sci. Data.*, **8**, 170, https://doi.org/10.1038/s41597-021-00964-1.

Ouyang, W., K. Lawson, D. Feng, L. Ye, C. Zhang, and C. Shen, 2021: Continental-scale streamflow modeling of basins with reservoirs: Towards a coherent deep-learning-based strategy. *J. Hydrol.*, **599**, 126455, https://doi.org/10.1016/j.jhydrol.2021.126455.

Painter, T. H., and Coauthors, 2016: The airborne snow observatory: Fusion of scanning lidar, imaging spectrometer, and physically-based modeling for mapping snow water equivalent and snow albedo. *Remote Sens. Environ.*, **184**, 139–152, https://doi.org/10.1016/j.rse.2016.06.018.

Paszke, A., and Coauthors, 2019: PyTorch: An imperative style, high-performance deep learning library. *NIPS'19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, H. Wallach et al., Eds., Curran Associates, Inc., 8024–8035, https://dl.acm.org/doi/10.5555/3454287.3455008.

Qin, Y., and Coauthors, 2020: Agricultural risks from changing snowmelt. *Nat. Climate Change*, **10**, 459–465, https://doi.org/10.1038/s41558-020-0746-8.

Rahmani, F., K. Lawson, W. Ouyang, A. Appling, S. Oliver, and C. Shen, 2021a: Exploring the exceptional performance of a deep learning stream temperature model and the value of streamflow data. *Environ. Res. Lett.*, **16**, 024025, https://doi.org/10.1088/1748-9326/abd501.

——, C. Shen, S. Oliver, K. Lawson, and A. Appling, 2021b: Deep learning approaches for improving prediction of daily stream temperature in data-scarce, unmonitored, and dammed basins. *Hydrol. Processes*, **35**, e14400, https://doi.org/10.1002/hyp.14400.

Rhoades, A. M., P. A. Ullrich, C. M. Zarzycki, H. Johansen, S. A. Margulis, H. Morrison, Z. Xu, and W. D. Collins, 2018: Sensitivity of mountain hydroclimate simulations in variable-resolution CESM to microphysics and horizontal resolution. *J. Adv. Model. Earth Syst.*, **10**, 1357–1380, https://doi.org/10.1029/2018MS001326.

Serreze, M. C., M. P. Clark, R. L. Armstrong, D. A. McGinnis, and R. S. Pulwarty, 1999: Characteristics of the western United States snowpack from Snowpack Telemetry (SNOTEL) data. *Water Resour. Res.*, **35**, 2145–2160, https://doi.org/10.1029/1999WR900090.

Seyednasrollah, B., and M. Kumar, 2013: Effects of tree morphometry on net snow cover radiation on forest floor for varying vegetation densities. *J. Geophys. Res. Atmos.*, **118**, 12 508–12 521, https://doi.org/10.1002/2012JD019378.

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

——, and K. Lawson, 2021: Applications of deep learning in hydrology. *Deep Learning for the Earth Sciences*, John Wiley and Sons, 283–297, https://doi.org/10.1002/9781119646181.ch19.

——, and Coauthors, 2023: Differentiable modelling to unify machine learning and physical models for geosciences. *Nat. Rev. Earth Environ.*, **4**, 552–567, https://doi.org/10.1038/s43017-023-00450-9.

Siirila-Woodburn, E. R., and Coauthors, 2021: A low-to-no snow future and its impacts on water resources in the western United States. *Nat. Rev. Earth Environ.*, **2**, 800–819, https://doi.org/10.1038/s43017-021-00219-y.

Slater, A. G., and M. P. Clark, 2006: Snow data assimilation via an ensemble Kalman filter. *J. Hydrometeor.*, **7**, 478–493, https://doi.org/10.1175/JHM505.1.

Smyth, E. J., M. S. Raleigh, and E. E. Small, 2022: The challenges of simulating SWE beneath forest canopies are reduced by data assimilation of snow depth. *Water Resour. Res.*, **58**, e2021WR030563, https://doi.org/10.1029/2021WR030563.

Stillinger, T., K. Rittger, M. S. Raleigh, A. Michell, R. E. Davis, and E. H. Bair, 2023: Landsat, MODIS, and VIIRS snow cover mapping algorithm performance as validated by airborne lidar datasets. *Cryosphere*, **17**, 567–590, https://doi.org/10.5194/tc-17-567-2023.

Swenson, S. C., and D. M. Lawrence, 2012: A new fractional snow-covered area parameterization for the Community Land Model and its effect on the surface energy balance. *J. Geophys. Res.*, **117**, D21107, https://doi.org/10.1029/2012JD018178.

Tsai, W.-P., D. Feng, M. Pan, H. Beck, K. Lawson, Y. Yang, J. Liu, and C. Shen, 2021: From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling. *Nat. Commun.*, **12**, 5988, https://doi.org/10.1038/s41467-021-26107-z.

Ullrich, P. A., Z. Xu, A. M. Rhoades, M. D. Dettinger, J. F. Mount, A. D. Jones, and P. Vahmani, 2018: California's drought of the future: A midcentury recreation of the exceptional conditions of 2012–2017. *Earth's Future*, **6**, 1568–1587, https://doi.org/10.1029/2018EF001007.

USDA, 2022: Snow program overview. USDA NRCS NWCC, accessed 8 April 2022, https://www.nrcs.usda.gov/wps/portal/wcc/home/aboutUs/snowProgramOverview/.

Vano, J. A., 2020: Implications of losing snowpack. *Nat. Climate Change*, **10**, 388–390, https://doi.org/10.1038/s41558-020-0769-1.

Venäläinen, P., K. Luojus, J. Lemmetyinen, J. Pulliainen, M. Moisander, and M. Takala, 2021: Impact of dynamic snow density on GlobSnow snow water equivalent retrieval accuracy. *Cryosphere*, **15**, 2969–2981, https://doi.org/10.5194/tc-15-2969-2021.

Vrugt, J. A., H. V. Gupta, B. Nualláin, and W. Bouten, 2006: Real-time data assimilation for operational ensemble streamflow forecasting. *J. Hydrometeor.*, **7**, 548–565, https://doi.org/10.1175/JHM504.1.

Winstral, A., and D. Marks, 2002: Simulating wind fields and snow redistribution using terrain-based parameters to model snow accumulation and melt over a semi-arid mountain catchment. *Hydrol. Processes*, **16**, 3585–3603, https://doi.org/10.1002/hyp.1238.

Xiang, Z., J. Yan, and, I. Demir, 2020: A rainfall-runoff model with LSTM-based sequence-to-sequence learning. *Water Resour. Res.*, **56**, e2019WR025326, https://doi.org/10.1029/2019WR025326.

Zeiler, M. D., 2012: ADADELTA: An adaptive learning rate method. arXiv, 1212.5701v1, https://doi.org/10.48550/arXiv.1212.5701.

Zeng, X., P. Broxton, and N. Dawson, 2018: Snowpack change from 1982 to 2016 over conterminous United States. *Geophys. Res. Lett.*, **45**, 12 940–12 947, https://doi.org/10.1029/2018GL079621.

Zhi, W., D. Feng, W.-P. Tsai, G. Sterle, A. Harpold, C. Shen, and L. Li, 2021: From hydrometeorology to river water quality: Can a deep learning model predict dissolved oxygen at the continental scale? *Environ. Sci. Technol.*, **55**, 2357–2368, https://doi.org/10.1021/acs.est.0c06783.

——, W. Ouyang, C. Shen, and L. Li, 2023: Temperature outweighs light and flow as the predominant driver of dissolved oxygen in US rivers. *Nat. Water*, **1**, 249–260, https://doi.org/10.1038/s44221-023-00038-z.